Uncertainty in Machine Learning Models

Oparah, F. C

2023

2023 Oparah F.C M.Sc

Uncertainty in Machine Learning Models

By

Oparah, F. C

Dissertation submitted to the University of Derby in partial fulfilment of the requirements for the award of Master of Science in Big Data Analytics

2022 - 2023

Abstract

Project Aim: As Artificial Intelligence and machine learning (ML) take centre stage in many human endeavours, the need to provide trustworthy predictions underpinned this study. Thus, the effective application of ML models to real-world phenomena requires the ability to quantify and limit the uncertainties in model predictions. Consequently, this study investigates the challenge of uncertainty quantification in machine learning regression and classification problems, with a view to providing the basis for reliability in ML model predictions.

Observed condition: In the traditional statistical analysis, uncertainties in the estimated quantities are accounted for with confidence interval construction. However, the equivalence of such uncertainty quantification in machine learning models has been sparingly studied.

Methods: The Frequentist method was adopted to examine uncertainty in the traditional statistical system and demonstrate confidence interval construction in making inference about population parameters. Conversely, the ensemble method of uncertainty quantification in ML and bootstrap sampling were also adopted by this work to examine a supervised ML system. Hence, the LightGBM and Random Forest (RF) algorithms were employed in the regression problem, while Decision Tree (DT) and LightGBM were adopted for the classification problem.

Key findings: Regression results from the Frequentist method showed moderately high model goodness of fit given by an R-squared of 0.741, while confidence intervals for the parameter estimates were constructed at the 95% and 90% levels. Point estimation results for the ML regressors suggest RF performed better than LightGBM with R-squared of 0.862 and 0.828 respectively. Evidence form the Root Mean Square Error (RMSE) and the Mean Square Error (MSE) also suggests better prediction for the ML models over the traditional model. The Prediction Interval Coverage Probability (PICP) and Mean Interval Prediction Width (MIPW) which measure uncertainty showed that RF with a PICP of 0.914 produced better coverage. Furthermore, the classification results show LightGBM with an accuracy rate of 0.81 and Area Under the Curve (AUC) of 0.86 was a better model than DT. Similarly, the variability in the classification results captured by the Standard Deviation (SD) of bootstrap samples showed a good measure of uncertainty. The results further revealed that LightGBM has less variability in its prediction with an average SD of 0.15.

Conclusion: Although ML model presents intricacies with varying uncertainty evaluation methods, it produces better uncertainty quantification as it does not mainly rely on confidence intervals like the traditional statistical model.

Acknowledgements

My sincere gratitude goes to my supervisor, Dr. Lee Barnby who in his busy schedules gave me constructive guidance and support throughout this study. Your direction, feedbacks and tutelage contributed to deepen the quality of this work.

I appreciate all my lecturers who through their efforts, I have been impacted with academic brilliance throughout my study. I equally thank other staff members of the University and my colleagues who have contributed immensely to make my study worthwhile.

Table of Contents

List of Tables

List of Figures

1. Introduction

1.1 Background of study

The current ubiquitous nature of Artificial intelligence (AI) and Machine Learning (ML) have altered the manner businesses and private endeavours are conducted globally. AI and ML make a sharp deviation from the brick and mortar approaches of the traditional society to a technologically embedded domain, where ML modeling has provided promising outcomes. However, it is believed that even high quality machine learning algorithms are anticipated to produce imperfect results. Such imperfections are sometimes systematically captured in probable intervals within which actual evaluations may lie, technically known as uncertainty modeling (Konig, Hoos and Rijn, 2020). While predictive ML algorithm possess ample promises and are increasingly adopted in multiplicity of intricate tasks, the outcomes are not completely reliable owing to the challenges presented by uncertainty (Butvinik, 2022).

ML models do not learn to produce real functions, but mainly empirical approximations of the real relationship between the model input and its outcome. Given these approximations, the accuracy of the model's output is limited by the model's parameters, input variables, hyperparameters and available data points (Imerit, 2022). In a supervised ML scenario, which is a case of learning input-output relationships from empirical data, the problem is either regression, for continuous outputs, or classification, for discrete outputs. In these scenarios, ML models generate optimal solutions based on their training data and are typically used for several forms of inferential decision making. In several applications, such inferences are inherently uncertain. Thus, ML models predictions which are used in diverse areas such as medical diagnoses, investment options, weather forecasting, security surveillance, etc contend with uncertainties, which ultimately affect decisions making. Therefore, it becomes imperative to evaluate the reliability of these models before deploying them, since the resulting predictions are exposed to noise and model inference errors (Antoran *et al,* 2021).

Globally, systems that rely on models belonging to the domains of ML and AI are becoming increasingly indispensible. Thus, the effective application of ML models to predict real-world phenomena require the ability to quantify and limit the uncertainties in model predictions, through the provision of valid and accurate prediction quantities. Thus, besides making a prediction, we must know how confident we are in this prediction. Both uncertainty estimation and interpretability are key factors for trustworthy machine learning systems (Abdar *et al*, 2021).

The principles of uncertainty play an important role in both the traditional statistical modeling and ML. In a traditional statistical analysis, estimated uncertainties in the analysed quantities are usually propagated to the final result giving a so-called error-band. Similarly, statistical models of data produce estimates with confidence intervals. For instance, in the traditional Least Squares (LS) regression models, the output produces the model parameter estimates together with their uncertainties, characteristically by way of probability distribution. The distribution is usually described by the mean and variance (or standard deviation). This is then followed by the construction of the confidence intervals with the probability derived from analytical procedures. This systematic methodology is based on the belief that the mean and variance are adequate in describing the distribution close to normal distribution (Edanz, 2023)

There are many sources of uncertainty in ML modeling, including variance in the specific data values, training data collection, imperfections in model parameters (model inadequacy, bias or discrepancy) and in the imperfect model structures developed from such data. ML algorithms are typically fixated on model training and validation of models, with minimal emphasis on the examination of statistical properties of the estimated quantities, including the estimation of the uncertainties associated with model output. ML algorithms modify model structures and parameters by the use of sample training data. In the course of model training, the selected feature criterion is optimized. This is followed by model validation. The validation stage involves additional set of the data, usually not employed in the model training. The validation could also entail a cross validation done recurrently with numerous data divided into training and testing sets. By doing this, the model's generalization capability is rationalized. Although cross-validation techniques could offer extra tools for numeric evaluation of uncertainty, it imposes increasing needs for a larger volume of computation than used for model estimation (Solomatine and Shrestha, 2009; Arkov, 2022). Hence, this imposes a drawback on uncertainty estimation under ML. However, with stable number of model parameters/hyperparameters enabled by the availability of more computing power, there has been noticeable introduction of new methods in uncertainty estimation. In general, the need for uncertainty quantification (UQ) underpins many key decisions; and model predictions without UQ usually lack trustworthiness (Abdar *et al.*, 2021).

Given that uncertainties are intrinsically connected with decision making processes, it becomes pertinent to seek answers to the following questions: What it the equivalent of uncertainty for a model result obtained through machine learning? This study seeks to find answers to this by exploring with supervised problems, using real world data from secondary sources.

1.2 Aim and Objectives of study

1.2.1 Aim of Study

The aim of this study is to investigate how uncertainty is quantified in machine learning models so as to provide reliability and trust to machine learning model predictions.

1.2.2 Objectives of Study

The objectives of the study are to:

i. Investigate uncertainty in the traditional regression model using the Ordinary Least square method.

ii. Estimate uncertainty in Machine learning using regression models

iii. Examine uncertainty in Machine learning using classification models

iv. Make a comparative analysis of uncertainty estimates under the traditional statistical methods and ML methods.

v. Provide recommendations on the effective methods of incorporating uncertainty quantification into ML models so as to provide reliability on ML predictions.

2. Literature Review

2.1 Introduction

To reinforce the investigation of uncertainty in ML, this section examined the nature of uncertainty, the nature of machine learning and various key concepts applicable to this study. This will create a thorough understanding of the subject matter. Furthermore, other related studies were examined in this section.

2.2 Meaning of Uncertainty

In general terms, uncertainty is construed as “what is not exactly known” or it could be said to be a condition which comprises unknown or imperfect information. In statistical parlance, uncertainty refers to how estimates might differ from the ‘true value” or the measurement of a doubt in the occurrence of an event. When a measurement b′ is performed on a particular quantity intended to be measured (measurand) b ϵ (-∞, ∞) with a true value b*. Many times b′ and b* differ by an error term e= |b′ - b*|. The error is the summation of various disturbances such as inaccuracies or missing observations. Therefore, the real value b* is required to be able to compute such a deviation. The uncertainty is either known, making the measurand *uncertainty-aware*, or unknown, resulting in an *uncertain* measurand (Hariri, Fredericks and Bowers, 2019).

Uncertainty can be categorized differently, for instance by considering its causes. Thus, various classifications of uncertainty exist in literature, which provide varying perspectives on uncertainty, such as aleatoric, epistemic, irreducible, reducible, predictive, and statistical, etc. Hence, there are several types of uncertainty in data analytics that could negatively affect the accuracy and reliability of the results. For instance, if the training data is incomplete, biased or acquired through improper sampling technique, the learning algorithm adopting such corrupted data is very likely to produce erroneous results. Thus, it is very important to introduce reliable techniques to quantify and handle uncertainty. To handle the many types of uncertainty that exist, several theories and techniques have been developed for the required modeling and resolution (Hariri, Fredericks and Bowers, 2019).

2.3 Taxonomy of Uncertainty

Different classifications and categorization of uncertainties relevant to this study are identified. Uncertainty emanates from various sources that could be sub-divided into types of uncertainty as shown in figure 1 below:

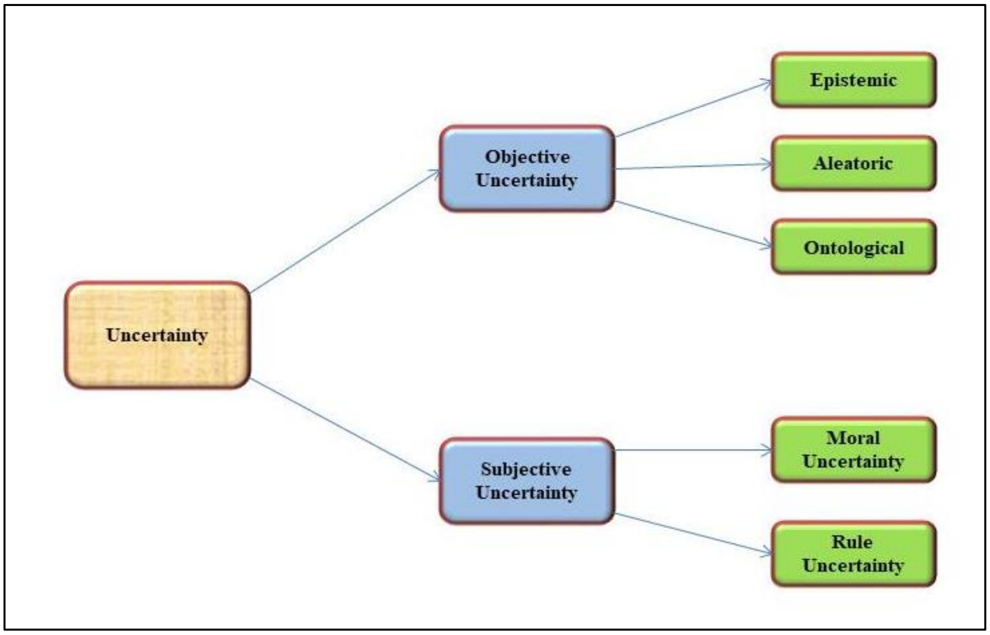


Figure 1: Taxonomy of Uncertainty (adopted and modified from Souza *et al.*, 2019).

2.3.1 Objective Uncertainty vs Subjective Uncertainty

Objective uncertainty refers to uncertainty that can be measured from the available information. Barthelmé and Mamassian (2009) believe uncertainty depends of the quality of available information, hence employing a Bayesian model, they showed that objective uncertainty is connected to subjective uncertainty. Campos, Neves and Souza (2007) identified objective uncertainty as corresponding to the variability that occurs from the stochastic nature of an environment, heterogeneity of the materials, space and time variations, or other types of dissimilarities amongst mechanisms or individuals. This difference emphasizes its relationship with the random features of games of chance, hence are stochastic in nature or "Irreducible". This is also called Uncertainty Type 1.

On the other hand, Subjective Uncertainty refers to the uncertainty that arises from scientific unawareness, lack of measurement, absence of confirmation or any other type of knowledge dearth. It is also called Uncertainty Type 2, or Reducible Uncertainty. It is believed to be able to reduce through additional empirical analysis. Objective uncertainty has well been explored through classical probabilities (Barthelmé and Mamassian (2009).

2.3.2 Aleatoric Vs Epistemic Uncertainty

In approximate notation, Aleatoric uncertainty denotes the idea of randomness, or the variability in the result of an experiment owing to random effects innate in natural phenomena. It connotes the breadth of noise intrinsically found in the data which cannot be reduced by additional procedures. This irreducible uncertainty in data does not relate to the characteristic of the model, but rather is an intrinsic characteristic of the data distribution (Abdar, 2021). Aleatoric uncertainty emanates from corruptions of the input data and regardless of the quality of the input data, this uncertainty cannot be evaded.

Hüllermeier and Waegeman (2021) believe that an ideal instance of aleatoric uncertainty is tossing a coin. The process of data generation in this experiment has a stochastic element which cannot be reduced by the introduction of any extra information. Hence, the model can only generate random outcomes. Accordingly, it is assumed that the best model of this arrangement only has the capacity to produce two possible outcomes, in form of probabilities-(heads and tails). There is no certain outcome. Thus, Aleatory uncertainty can be said to be the unpredictability of an even that can only be fully described using probability.

The quantification of Aleatory uncertainty can be made in the form of probability distributions. It comprises:

i. Measurement uncertainty in which it is impossible to estimate input and output variables with complete precision. Thus, the entire measurements are susceptible to certain forms of inaccuracy.
ii. Sampling uncertainty which is usually experienced when investigating a random sample whose population is large. The sample may involve effects such as spatial, temporal momentary, exaggerated or miss effects. This variance is generally captured by the stochastic error term.

Conversely, epistemic uncertainty denotes uncertainty resulting from absence of knowledge about the structure and parameters of models. In cases where there is abundant collection of data, the accompanying information might be poor. It describes the reliability of the model in its representation of the data—exclusive of aleatoric uncertainties (Caldeira and Nord, 2021). In other words, it denotes the ignorance in decision making, in place of any underlying random or stochastic phenomenon. In contrast to uncertainty emanating from randomness, uncertainty resulting from ignorance could decrease with the introduction of extra information. Thus, epistemic uncertainty rises when making predictions distant from known data and declines as the training data size increases. It also comprises uncertainties which arise when the model parameters are specified inappropriately with vague knowledge or lack of direct measures. Furthermore, uncertainties in the structures of the model due to bias and model discrepancies constitute epistemic uncertainty. Errors introduced in the modeling process due to the desire to reduce computational cost and model complexities also constitute epistemic uncertainty (Hüllermeier and Waegeman, 2021)

Figure 2 below is a diagram showing the difference between aleatoric and epistemic uncertainties. The points on the diagram symbolize the existing data points. Aleatoric uncertainty covers various amount of noise inherent in the data, while epistemic uncertainty mirrors the ignorance gap resulting from absence of data.

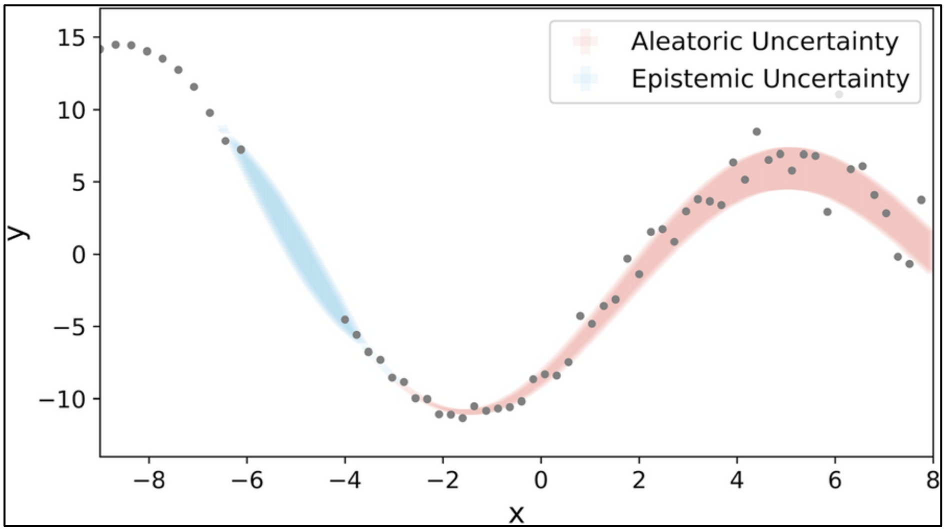


Figure 2: Aleatoric and Epistemic Uncertainty

Aleatoric and epistemic uncertainty can also exist concurrently in one period. For instance, when model parameters exhibit aleatoric uncertainty in an experiment, and those parameters are input to another experimental model. If thereafter for the purpose of uncertainty quantification an alternate model, e.g. a Gaussian process is learned from computer experiments, this alternate model displays epistemic uncertainty that interacts with the aleatoric uncertainty of the initial model parameters. Such uncertainty cannot exclusively be categorized as aleatoric or epistemic anymore but is more broadly classified as inferential uncertainty. In real-life situations, both kinds of uncertainties are present (Butvinik, 2022; Indrayan, 2020).

2.3.3 Ontological uncertainty

Ontological uncertainty is a complete lack of knowledge and understanding about the entire modeling process. It emanates from unconscious use of incongruous procedures or beliefs in the modeling process. Ontological uncertainty is mainly unrecognized, unquantifiable and incorporates data, techniques or models. It comprises what is known as Semantic uncertainty, which occurs when different connotations are attributed to the same terms, situations or actions. It also occurs when procedural descriptions are not clear or are inappropriate for a complete rational understanding. It also includes interpretational uncertainty which occurs when information encoded in data or models follows inconsistent methodology which lacks clarity. This uncertainty is very distinct in that it cannot be subjected to probability distribution. The attached meaning is unknown and unfamiliar and therefore it is impossible to predict the value that will emerge (Lane and Maxfield, 2004; Schmitt, 2023).

2.3.4 Moral Vs Rule Uncertainty

Moral uncertainty is uncertainty about what is well considered morally right to do. It is uncertainty about moral or evaluative matters. It also involves how to act given that one is morally uncertain. It is triggered by absence of appropriate moral rules. Hence, decision-makers will have no other option that to rely on more general moral rules and apply the same to infer direction in the particular circumstances under consideration. Regrettably, suppositions informed by general moral guidelines usually produce poor satisfaction to the decision-maker (Macaskill, Bykvist, and Ord, 2020). On the other hand, Rule uncertainty is uncertainty relating to decisions informed by intuition. That is, uncertainty based on moral rules. In particular circumstances, decisions can only be made by relying on our intuition instead of knowledge. This infers that such actions are based on general moral convictions, formed beforehand.

2.4 Uncertainty Quantification (UQ)

ML models can produce confident but incorrect predictions. However, to solve this problem, ML model developers adopt several techniques to reliably measure ML uncertainty. This is normally done on controlled benchmarks immediately the model is trained (Sanchez *et al.*, 2022). UQ could be defined as the process of assigning numerical values to uncertainties connected with model estimation of true, physical quantities of interest, with the aim of recognizing all pertinent sources of uncertainty and calculating the influences of particular sources to the general uncertainty (National Research Council, 2012). It could also be seen as an arrangement of statistical tools that define the uncertainties accompanying a particular model, with the aim of reflecting all possible and reasonable uncertainty sources so as to correctly assess an inclusive uncertainty (Abdar *et al.*, 2021).

UQ methods are very important in reducing the impact of uncertainties in the course of modeling as well as in decision making processes. Arguably, UQ presently underpins several key decisions; hence reliable predictions must be accompanied by UQ. The uncertainty associated with the result of a model could be specified by defining a range of values that are likely to encompass the true value of a particular variable. Within this notion, the best model is the one that is able to appropriately optimize a variable while minimizing the breadth of the uncertainty bands (Freni and Mannina, 2010). The estimation of the uncertainty associated with a predictive model is becoming a highly fundamental prerequisite in the interpretability of predictions. The goal of uncertainty quantification (UQ) is the description, identification, management and reduction of uncertainties in models especially, computer-based models and engineering systems. The adoption of UQ, makes it is possible to calculate the likelihood of likely certain results, if certain aspects of the model of interest are either not fully known, subject to stochastic variations, or only incomplete information is available for some parts of the system(Frau *et al.*, 2021).

The procedure for recognizing and classifying the different sources of uncertainty is intricate and model based (Volodina and Challenor, 2021). UQ methods are valuable to minimize the impact of uncertainties on decisions based on model outputs. Large and complex computer-based models require a high degree of statistical proficiency and specific algorithms. Likewise, many of the traditional approaches for UQ require high degree of model evaluations which make it uninspiring. However, several approaches have been adduced in literature to handle this such as the Ensemble methods, Gaussian process, Monte Carlo drop-out, Bayesian neural networks and the Quantile regression. The merits and shortcomings of the each method gave rise to the development of a competing approach (Choubineh *et al.*, 2023).

2.4.1 UQ in the traditional statistical system

In the traditional statistical system of analysis, uncertainty is quantified by a probability distribution which is a function of the state of information regarding the likelihood of what true value of the uncertain phenomenon is. Generally, there are two famous methods to statistical inference and hypothesis testing, that is the Frequentist (or traditional) and the Bayesian methods. Accordingly, there are two contending approaches to uncertainty evaluation in the traditional system (Giaquinto *et al.*, 2014). In dealing with a core problem in uncertainty evaluation which is propagation, it is usually the practice that estimated uncertainties in the analysed quantities are propagated to the final result with an error-band. Besides, models of data produce estimates with confidence intervals. Generally, the Frequentist method of inference believes events are founded on frequencies, while the Bayesian inference based its principles on prior knowledge (Hariri *et al.*, 2019).

i. The Frequentist approach

The Frequentists posit that a numerical value of a probability is the limit of relative frequency in a large number of trials. The Frequentists's statistical inference investigates whether an event occurs and they treat probabilities as equivalent to "frequencies". In the scientific framework, an experiment is assumed to be an infinite sequence of possible repetitions of a particular trial. It is understood to contain the drawing of samples from a distribution of a population of potential observations. This school of thought was advanced by great statisticians such as Jerzy Neyman, Ronald Fisher, and Egon Pearson during the early 1900s. They adopt tools such as p-values, confidence intervals, and statistical tests of significance. It was the belief of the Frequentists that probability depends on the result of an experiment if it was repeated infinitely. For example, if a coin is flipped infinite times and half of the time the outcome is heads, then it is accurate to infer that the probability of getting heads (or tails) is 50% (Giaquinto *et al.*, 2014)

Among the key estimates in the Frequentist method is the *p*-value. It validates hypothesis and denotes the probability of obtaining a result as extreme as obtained if the experiment or study was repeated, assuming the null hypothesis was true. The *p*-value varies based on the underlying distribution. Usually, a level of significance is indicated beforehand and indicates how small the *p*-value should be to reject the null hypothesis. This level of significance can differ depending on the practice or norm. Similarly, the Frequentist confidence interval shows how sure it is that if the experiment was repeated, the outcome will be in a given range typically set at 95% (Edanz, 2013).

The Ordinary Least Squares (OLS) method of regression analysis is an example of a popular method in the traditional statistical epoch. Linear regression in general is one of the most widely used statistical tools in scientific applications. The typical practice is to consider several predictor variables in hoping that a few will stand out as being important in explaining variations in the response/dependent variable. Consequently, there is large uncertainty in the underlying model.

The Frequentist method of handing this issue is to use the data in selecting a particular model. It employs a variety of criteria such as Akaike information criterion (AIC), the Bayesian information criterion (BIC), Root Mean Squared Error (RMSE), Residual Standard Error (RSE), etc. However, these Frequentist methods were unable to provide quantification of uncertainty about the true model different from confidence intervals. In other words, they did not provide any process to make a precise conclusion that one model being investigated is more plausible in being the true model than another model (Martin *et al.*, 2018; Ivanov, 2020).

ii. Bayesian method

The Bayesian method is based on Bayes' Theorem and was advanced by Thomas Bayes (Bayes, 1763). Its form is of a statistical formula that stipulates the probability of the occurrence of an event in view of a prior knowledge of circumstances connected to that event. This theorem projects in what way existing beliefs should be updated with data in order to derive new beliefs. It is based on the Bayesian interpretation of probability in which probability expresses a degree of information (knowledge) about an event. This is different from the Frequentist interpretation which views probability as the limit of the relative frequency of an event after many trials (Giaquinto *et al.*, 2014). Notwithstanding, the conceptual straightforwardness of the Bayesian method, certain practical difficulties must be overcome, such as specification of prior values. In specific terms, real prior information is hardly available and cannot be substituted with improper default priors.

The Bayesian statistics is underpinned by the Bayes' theorem**,** and the sum and product rules of probability to update priors *p(y)* when more evidence *p(x|y)* is gotten, giving rise to a posterior probability distribution *p(y|x)*. The Bayesian theorem is presented in equation (1) as given below:

$$p(y|x) = \frac{p(x|y)p(y)}{p(x)} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (1)$$

Where:

p(*y*) = the probability of y occurring,

p(*x*) = the probability of event x occurring,

P(*y*/x)= probability of y given event x

P(x/y) = probability of x given event y.

In the Bayesian scenario, y equals the parameters and x is the data. From equation (1), p(*y*|*x)* is known as the *posterior*. P(*y*) is the prior, implying the probability assumed by parameters before the experiment. P(*x*|*y*) is the likelihood, which is the probability of the data given the parameters. The sum of all possible arrangements of y of the distribution *p(x,y)* = *p(x*|*y)p(y)* in (1) gives rise to the marginal distribution of *x*, which is *p(x).* Finally, *p(x)* is the probability of the data, and it is subject to difficult computation which is a low side of the Bayesian approach.

In principle, the linear regression $\mathbf{Y= \beta_0 + \beta_1X} + \boldsymbol{e}$ just as any parametric model can be expressed as a corresponding Bayesian model by stating priors over the parameters β_i. (Where: Y is the response variable, X is the explanatory variable, β_0 and β_1 are the model parameters, and e is the error term). The Bayesian method is able to generate a summary of the uncertainty among contending models; see, for instance, Clyde and George (2004). Beginning from a prior distribution (*p(y))* on the set of possible models and a set of conditional priors ($p(x|y)$) on the model-specific parameters, a posterior distribution ($p(y|x$) on the model space can be obtained via Markov chain Monte Carlo technique (Pevec and Kononenko, 2015).

In Bayesian analysis, prior distributions is assigned to the unknown parameters (β_0, β_1, and σ^2) based on our prior knowledge or beliefs. These prior distributions represent our initial uncertainty about the parameter values. Given the data (x, y), the goal is to update our prior beliefs and ascertain the posterior distribution of the parameters using Bayes' theorem. The posterior distribution denotes our updated knowledge about the parameters after considering the observed data. To perform the Bayesian analysis, we typically choose conjugate prior distributions that results in closed-form posterior distributions. Once we have the posterior distribution, we can derive various quantities of interest, such as point estimates (e.g., posterior mean, median, or mode) or credible intervals, which provide a range of plausible values for the parameters.

The Bayesian approach relies on certain hypotheses which have to be cautiously analysed. In general applications, such as the linear regression, the hypothesis is made regarding the normal distribution of the residuals between the model and observations assuming the zero mean and variance, σ^2 (Yang *et al.*, 2008).

2.4.2 UQ in Machine learning Algorithms

Typical ML and Deep learning techniques for regression and classification do not define model uncertainty (Choubineh *et al.*, 2023). In classification models, predictive probabilities obtained as outputs are often erroneously interpreted as model confidence (Gal and Ghahramani 2016). Furthermore, ML models cannot give information regarding the reliability of their prediction (black box models), despite their successes at solving real world problems (Abdar, 2021).

ML models are expected to produce calibrated uncertainty estimates. Poorly calibrated uncertainty estimates employed as input results in biased estimates of physical quantities.

ML algorithms present uncertainty evaluations as probability distributions (Chen *et al.,* 2022). These could be class probabilities for classification problems or full probability distribution in the cases of regression problems. Calibration denotes the ability of such probability estimates to match the frequency of occurrence of the target true population variable. ML models extract information from data and calibrate model structure/parameters using sample training data. During the model training, the predictive objective function is optimized while the test data serves as a validation tool. Validation and cross-validation of the model result in establishing the model's generalization ability.

Arguably, ML algorithms follow many statistical techniques such as the Ordinary Least Squares (OLS). The latter adopts the Gaussian Process and yields parameter estimates with zero mean and constant variance, in which uncertainty is captured by specifying probabilities as well as confidence intervals for the true parameters. However, the current ML algorithms generally concentrates on training and validation of models, without giving adequate attention to the statistical properties of the estimates obtained, including uncertainty (Arkov, 2022). Cross-validation procedures offer further tools for numeric assessment of uncertainty; however, it imposes increasing requirements for more complex computation than employed in estimating the model.

Generally in ML, predictive uncertainty associated with the output depends on both the epistemic uncertainty and aleatoric uncertainty, which are basically used to differentiate the sources of errors. However, while these terms are occasionally used interchangeably, they could differently apply in certain cases and model-building scenarios. Predictive uncertainties also depend on the divergence between the data domains. For instance, a deep neural network (DNN) trained exclusively on numerical cosmological simulations will have more uncertainties attributable to it whenever it is subjected to real telescopic observations. Such dataset-dependent uncertainties could be due to a domain transition (when the distribution of training data does not match with those of the real data) or due to inadequate domain information. Given the above scenario, uncertainty quantification must be specific to a particular approach in a problem under investigation.

In recent times, several applications of UQ algorithm have begun to transform the settings of AI and ML deterministic models to models involving more probabilistic genre. Different methods of UQ in machine learning could be identified. Abdar *et al.,* (2021) identified three main types of UQ: (i) ensemble techniques such as deep ensemble, deep ensemble Bayesian/Bayesian deep ensemble, etc (ii) Bayesian methods such as Monte Carlo (MC) dropout, Markov Chain Monte Carlo, ect, (iii) Sundry methods such as deep Gaussian Process (GP), Quantile regression, etc.

i. **The Gaussian processes:** A Gaussian Process (GP) is a non-parametric regression procedure which attempts to establish a distribution for the response variable over diverse possible functions consistent with the input data. It is an assemblage of random variables, any finite number of which has joint Gaussian distributions. The Gaussian distribution refers to vectors, while the GP is about functions (Rasmussen, 2003). GP is usually defined by its mean function m(x) and covariance function k(x, x).

   A regular generalization of the Gaussian distribution whose mean and covariance is a vector and matrix, respectively could be written as: $f \sim GP(m, k)$. This implies that the function (f) is distributed as a Gaussian processes (GP) with mean, m and covariance, (k). Basically, the argument x of the random function f(x) plays the role of index set. Meaning that for every input x there is an associated random variable f(x), which is the value of the (stochastic) function *f* (Rasmussen & Williams, 2006).

   The GP regression is usually employed to model relationships or variables that do not have spatial or temporal dependence. They take the values of the variables at various special locations as correlated and see the correlation coefficient as a flexible function of special separation. The GP regression learns the correlation behaviour from data collected at known and fixed locations; this knowledge could be applied for prediction and interpolation in any subject area. Consequently, the interpolated value of the variable at any given location is estimated as a weighted average of the measured data subject to the weights being dependent on the special correlation performance. In GP regression the uncertainties associated with predicted values are automatically accounted for. The correlation structure in data is recognized by GPs through σ, the hyperparameters that define the covariance kernel. To make predictions from GPs, the training data provides learning for these hyperparameters by minimizing the log of the marginalized posterior (Thompson *et al.*, 2021).

Given that the Gaussian distribution is a probability distribution over a finite set of random variables $y, y \sim N(\mu, V)$, with a mean (μ) which is a vector and covariance matrix (V), the GP is a probability distribution with respect to functions y(x), $y(x){\sim}GP(\mu(x), V(x, x'))$

Where:
$y(x)$= the mean function
$V(x, x')$ is the covariance kernel.

An important advantage of GP is that it yields the expected value of the posterior distribution as well as the attendant variance which could be adopted to measure uncertainty. The choice of the kernel when constructing the GP regressor is very critical since the assumptions of prior information made about the function to be learned is embedded. In most cases the kernel is usually considered as a hyperparameter during the training of the kernel and thus different combinations are explored in practice (Frau *et al,* 2021)

ii. Monte Carlo drop-out

Monte Carlo-based techniques investigate the propagation of uncertainty of parameters to the probability distribution function (PDF) of the output (Solomatine and Shrestha, 2009). Monte Carlo Dropout is a form of advanced neural network (NN) method that adopts dropout regularization in order to create more trustworthy predictions and produce prediction uncertainties. It extends further than the traditional usage of dropout in model training and encompasses the inference stage. By adopting dropout for modeling inference, Monte Carlo Dropout produces various predictions for a single input, giving rise to more accurate quantification of uncertainty in the model's predictions (Rahman, 2023).

MC Dropout randomly drops certain units of NN at a specific probability in the course of a forward pass, over the network to stop them from co-tuning excessively. Ordinarily, these dropout layers are deactivated after training to avoid interference with the forward pass on a new data point. Dropout is applied both during training and inference. During inference, repetitive estimations with the same input x give varying results. It usually use the mean as the prediction while it use the standard deviation as the uncertainty.

The MC dropout procedure offers a scalable manner to learn a predictive distribution. MC dropout functions by randomly switching off neurons in a NN, which normalizes the network. Every dropout configuration relates to a different sample from the approximate parametric posterior distribution. Inferring predictive distributions with Bayesian NN is among the most widely used methods of estimating uncertainty. A predictive distribution (D) could be represented by:

$p(y|x, D)$

$D = (x_i, y_i)_{i=1}^{N}$ ………………………………………………………… (2)

Where:

y= target

x= input

N= several training examples

When a predictive distribution is obtained, through inspection of the variance and uncertainty could be uncovered. A predictive distribution could be learned over its distribution or functions. This could also be done through its parameters, i.e. parametric posterior distribution: $p(\theta|D)$.

$$\theta_t \sim q(\theta|D)$$

Where θ relates to a dropout configuration, or a simulation ~, sampled from the approximate parametric posterior: $p(\theta|D)$. Sampling from the approximate posterior $q(\theta|D)$ enables Monte Carlo integration of the model's likelihood, which uncovers the predictive distribution.

Modeling uncertainty with Monte Carlo dropout is implemented by running several forward passes through the model with varying dropout masks on each occasion. Given a trained NN model with dropout *fnn*, to obtain the uncertainty for one sample *x*, the predictions of *T* inferences are collected with different dropout masks. In this instance f_{nn}^{di} represents the model with dropout d_i. Hence a sample of conceivable model results for sample *x* could be obtained as follows: $f_{nn}^{d0}(x), \dots, f_{nn}^{dT}(x)$.

By calculating the mean and the variance of this sample an ensemble prediction is derived. This produces the mean of the models posterior distribution for the sample and an estimate of the uncertainty of the model in relation to *x*.

iii. **Ensemble methods:** Ensemble models are simply Meta machine learning models (ML algorithms that learn how best to combine the predictions from other ML algorithms) built from numerous smaller models. These different member models could all have similar or different architectures and be trained on smaller portions of the overall training dataset. Each associate model provides a prediction regarding what it considers the possible solution. Thereafter, all the predictions are combined to form a final prediction centered on some average or skewed sum of all the constituents. The uncertainty measure is obtained from how much these member models disagree with one another. (Dietterich, 2000; Hoffmann *et al.,* 2021)

In a regression problem, this could be demonstrated with the implementation of models built from a series of linear regressions, each trained on a different subset of the training data, or with a classification model. In a typical supervised training problem, a learning algorithm is assigned training specimens given by $\{(X_1,y_1),\ldots,(X_m, y_m)\}$ for certain unknown function y= f(X). The values of X_i are usually vectors given by $(x_{i,1},x_{i,2},\ldots,x_{i,n})$ whose constituents are discrete or real-values like age, colour, height, etc, which are the features of x_i. The values of y are normally chosen from a discrete class {1,…...,K} in a classification problem or from the real number line in a regression problem.

Given a set of *T* training instances, a learning algorithm produces a classifier output. The classifier is a hypothesis about the true function *f*. With new set of x values, it will predict the corresponding values of y and this will be represented as $h_1, \ldots h_k$. Given that an ensemble of classifiers is a set of classifiers whose different decisions are combined in some way naturally by weighted or average voting to classify new problems, a necessary and sufficient condition for an ensemble of classifiers to be more precise in capturing uncertainties than any of its members, is if they are accurate and diverse. A classifier is said to be accurate if it has an error rate that is better than guessing on new values of *X*. Two classifiers are not the same if they commit different errors on new data points.

iv. **Quantile Regression:** The Quantile regression assesses the estimated model residuals and all other causes of uncertainty which is different from other orthodox techniques like the Monte Carlo-based techniques. The latter characteristically considers only one source of uncertainty (Solomatine and Shrestha, 2009).

The Quantile regression (QR) is a linear statistical technique for estimating the quantiles conditional functions of model prediction based on possible causal relationships within the whole data set. It defines the conditional quantiles distribution as functions of observed covariates without making any assumptions about the nature of the data distribution. In this method, each quantile τ, has a linear relationship between the observed response variable(y) and predicted response variable $\hat{y}$ as show in equation 3 below:

$y= \alpha_\tau \hat{y} + b_\tau$ …………………………………………………..(3)

Where α_τ and b_τ are intercept and slope parameters of the model respectively.

2.5 Related studies

Several empirical studies have recognized the explanability of the traditional statistical models in quantifying uncertainty by the use of probability distribution and confidence intervals. It was equally observed that ML systems are by-products of standard statistical procedures, such as the Least Squares procedure. Usually, the Least Square procedure produces the model parameters estimates alongside their uncertainty estimates, usually in the form of the probability distribution (Der Kiureghian and Ditlevsen 2009; Arkov, 2022).

ML literature broadly made a distinction between aleatoric uncertainty and epistemic uncertainty. While aleatoric uncertainty refers to uncertainty inherent in data, epistemic uncertainty refers to uncertainty in the model (Kendall and Gal, 2017). However, orthodox approaches to probabilistic modeling, which are basically founded on bagging knowledge in terms of a single probability distribution, fail to separate these two fundamental different sources of uncertainty. Such distinction may not necessarily suffice, especially in the case of predictive modeling, as enunciated by the classical Bayesian decision theory (Hüllermeier and Waegeman, 2021).

Senge *et al.,* (2014) unequivocally identified the distinction between aleatoric and epistemic uncertainty. Consequently, they proposed a quantification of these uncertainties using a binary classification algorithm and similarly showed the usefulness of their method within the setting of medical decision making. Studies like Thompson *et al.,* (2021); Volodina and Challenor (2021); Kendall and Gal, (2017); and Chen *et al.,* (2022) also emphasized the distinction of these sources of uncertainty.

Inspired by the need to enthrone a trustworthy representation of uncertainty in ML, several studies have been carried out to examine the methods of uncertainty quantification and to propose bespoke alternative methods where limitations exist. Obvious recognition of uncertainties is not adequate. To have this embraced by decision makers, it should be properly quantified and communicated. Therefore, it becomes imperative to: (1) identify and understand the sources of uncertainty; (2) make a quantification of uncertainty; (3) assess the propagation of uncertainty through the models; and (4) investigate approaches to reduce uncertainty (Solomatine and Shrestha, 2009).

Hoffmann, Fortmeier and Elster (2021) adopted the ensemble learning to investigate uncertainty, in the area of computational optical measurements, which is a type of Deep technique. The reliability of the uncertainty measure was explored by methodically introducing out-of-distribution errors and noisy data in the model. The results strengthened the capacity of ensemble methods to make reliable predictions on super-dimensional real world data.

Konig, Hoos and Rijn (2020) recognized that uncertainty estimation and interpretability are central issues for trustworthy ML systems. Hence, they adopted a unique method for interpreting uncertainty approximations from differentiable probabilistic models, such as the Bayesian Neural Networks (BNNs). Their method called the Counterfactual Latent Uncertainty Explanations shows how to modify an input, while at the same time maintaining it on the data space, such that a BNN is now more assertive regarding the input’s prediction. The results shows that this technique outperforms baseline methods and allows analysts better understand the input sources of predictive uncertainty.

Arkov (2022) examined the uncertainty feature of mathematical modeling in ML. Regression model was employed in this study to investigate uncertainty in model parameters as well as in the output feature value predictions. The major stages in the orthodox Least Squares procedures to regression modeling were evaluated as well as uncertainty estimation. The study observes that model complexity in ML and severe nonlinearity are serious impediment to uncertainty quantification. Furthermore, the growing complexity of mathematical models constructed from real data is compounded by the increasing amount of training data requirement. Consequently, the study posits that the need to estimate uncertainty and assess the concomitant risks could be handled with non-parametric techniques based on cross-validation. The study concluded that box-and-whiskers plots and quantile measures of uncertainty offer a more flexible and favorable tool than standard deviation acceptable for Gaussian distribution.

Nagl, Nagl and Rösch (2022) in a study on "Quantifying uncertainty of machine learning methods for loss given default" observed that in credit risk milieu, when ML is applied to forecasting credit risk parameters, the methodologies have performed better than standard statistical techniques. However, the quantification of prediction uncertainty is not usually analyzed in ML. They noted that the quantification of uncertainty is very crucial as it deepens the transparency and resilience in risk management. Therefore, the study adopted the deep evidential regression technique which was applied to the loss given defaults (LGDs).This method was also useful in determining the uncertainty of regression tasks as well as in estimating the epistemic and the aleatoric uncertainty. The results further showed that aleatoric uncertainty is significantly larger than epistemic uncertainty. The results further suggest that aleatoric uncertainty is the main driver of the general uncertainty in LGD estimation

Tavazza, De Cost and Choudhary (2021) believes that while confidence intervals are commonly reported for ML models, the evaluation of the uncertainty on each prediction, is rarely obtainable. They compared three different approaches to obtaining specific uncertainty in ML models and tested them against ML physical properties. This was specifically investigated using the Quantile loss function, direct investigation of prediction intervals and using Gaussian Processes (GP). The results favoured direct modeling of individual uncertainties, given its simplicity and given that it generally, minimizes incorrect estimation of the predicted errors. They study maintained that the choice of hyperparameters is particularly crucial when developing GP models and GP models give a good estimate for prediction intervals.

Loquercio, Segu, and Scaramuzza (2020) opined that Neural Networks are not reliable when the input sample is corrupted by noise and the ability to automatically detect such inadequacy is a key to incorporate deep learning (DL) algorithm into robotics. They observed that existing methods for uncertainty estimation of Neural Networks require modifications to the network and optimization process, characteristically discard prior knowledge about the data, and are inclined make over-simplifying assumptions which underestimate uncertainty. In their study on "A General Framework for Uncertainty

Estimation in Deep Learning", proposed a framework based on Bayesian theory and Monte-Carlo sampling, to completely model the different sources of prediction uncertainty, incorporating prior information. They theoretically showed that this model captures uncertainty better than any other model.

Darling and Stracuzzi (2018) believed that although current ML accuracy based validation metrics show the performance of a classifier model, they do not indicate model predictive efficiency. Hence, the study developed some general theoretical foundations for quantifying uncertainty in supervised ML models by constructing probability distributions using an ensemble of classifiers. This was applied to a problem of detecting malicious websites. Thus, uncertainty quantification was used to assess the quality of the individual predictions made by supervised two-class classifiers. The results showed that the SD of prediction probability distributions has correlation with accuracy. However other measures such as highest density intervals were more informative.

3. Research Methodology

3.1 Introduction

The techniques for estimating uncertainty in both the traditional statistical system and ML which are adopted for this study are evaluated in this section. These methods were implemented on the selected dataset and the results were analysed accordingly.

3.2 Research Strategy

To investigate uncertainty, this study employs a supervised ML model. Consequently, linear regression model and binary classification model were adopted. For the traditional statistical system, the Frequentist method was adopted while the ensemble methods of uncertainty investigation were adopted for the ML modeling system. The choice of the Frequentist method is based on its adoption of the Ordinary Least Square (OLS) methods and the 'BLUE' properties of the OLS- Best Linear Unbiased Estimators (Gauss–Markov Theorem). It has also been shown in recent times that the contending method, the Bayesian method converges to the Frequentist method due to practical difficulties in prior specification. Real prior information is hardly available. Thus in the absence of genuine prior information, the model uncertainty assessments emanating from the corresponding posterior distribution are not guaranteed to be inferentially trustworthy (Martin *et al.,* 2018).

For the ML modeling, this study adopted a supervised learning technique following Xiaozhe et, al (2013); Omid et. At (2019); and Darling and Stracuzzi (2018). Consequently, regression and the classification models were adopted to investigate uncertainty. The LightGBM and Random Forest (RF) regressors were adopted for the regression problem. The choice of these models was because they incorporate the techniques of bagging and boosting in their approach to produce robust results. Similarly, the Decision Tree and the LightGBM classifiers were adopted for the classification problem. The choice of these classifiers was due to their impressive performance in problems involving heterogeneous features and high class imbalance. Training dataset to validation dataset ratio of 70%: 30% was adopted based on theoretical and empirical literature. The bootstrap method was incorporated in the classification problem to estimate uncertainty.

3.2.1 The Ordinary Least Squares (OLS) Method

The OLS is widely used method for estimating the parameters of a linear regression model in the Frequenst method. It seeks to find the parameter estimates that minimize the sum of the squared differences between the observed data points and the predicted values by the linear model. It does this without incorporating prior beliefs about the parameters, which is a characteristic of the Frequentist approach.

Ideally, the estimates of the linear regression model parameters are accompanied by their standard error (SE) or standard deviations written below the parameters as follows:

$$y = \underset{(\sigma_{\hat{b}0})}{\hat{b}_0} + \underset{(\sigma_{\hat{b}1})}{\hat{b}_1} x \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(4)$$

Where: y= response variable

x= independent variable

$\hat{b}_0$ = estimated intercept

$\hat{b}_1$ and estimated coefficient of the independent variable.

$\sigma_{\hat{b}0}$ = standard error of $\hat{b}_0$

$\sigma_{\hat{b}1}$ = standard error of $\hat{b}_1$

Usually, the modeling pipeline contains the stages as follows:

i) Estimating the regression model

ii) Making p-values, output feature point and interval values predictions.

iii) Plotting prediction and confidence intervals for the model output

iv) Estimating residuals and evaluating the residual plots to ensure they conform to the Gauss–Markov Theorem of normality, asymptotic consistency, efficiency, etc.

3.2.2 Bagging (bootstrap aggregating) and Boosting Techniques

Bagging involves training multiple instances of a single model on different subsets of the training data. Each subset is created by randomly sampling the training data with replacement. On the other hand, boosting, is an iterative technique that aims to correct the errors made by previous models in an adaptive manner. Each subsequent model focuses on the mistakes of its predecessors, attempting to improve upon them.

Random forest model adopted in the regression problem typically uses the bagging technique while the LightGBM which is a Gradient boosting Decision Tree adopts the boosting technique. The bootstrap technique was also applied to the Decision Tree and LightGBM models in the classification problem. The adoption of these techniques in this study is to produce robust results.

3.2.3 Bootstrap Sampling

Given a random sample X = (x_1, x_2, ..., x_n) from an unknown probability distribution *F*, the sampling distribution of random variable *R*(X, *F*) could be estimated based on the observed data, X. To implement this we sample from X with replacement to obtain a sampled set X* from the pool of observed data (Efron,1979). Boostrapping the dataset permits us to assess the range of possible output given the data and to create the variability required for building an active ensemble model (Kazmierczak *et al.,* 2022).

The bootstrap sampling was applied to the classification problem. After applying this to generate the sampled data sets and training models for each, we obtained the probability distributions of the candidate labels for each sample was obtained. The process is outlined as follows:

i. Obtain S bootstrap samples by sampling the data X = x_i, x_2,..., x_n, with replacement and obtain *n* data points.
ii. Designate the bootstrap samples x^{*1}, x^{*2}, ..., x^{*s}.
iii. For each of the S bootstrap samples, fit a classification model and estimate probability P_i for candidate labels y_i for each X*. The S values of p_i provide a probability distribution for each candidate label for each sample.

3.3 Uncertainty Quantification

3.3.1 Estimation under the traditional system

To measure uncertainty in the traditional statistical system, the Frequentist method was implemented to constructs a confidence interval for the parameter estimates. This could be regarded as the transposition of a hypothesis test to ascertain the set of plausible values for a parameter, given the observed data. The combination of the standard error, asymptotic normality and consistency, allows us to quantify uncertainty directly through confidence intervals. Under these conditions, the sampling distribution is approximately normal with mean (b_i)- the true parameter, and standard error ($\sigma_{\hat{b}i}$). A normal distribution has ≈ 95% of its mass within 1.96 standard deviations of the mean. Thus, the 95% confidence interval for the parameter estimates implies that the parameter estimates will lie within 1.96 standard error of the true value of b_i. Put differently, $\hat{b}_i - 1.96\ (\sigma_{\hat{b}1}) \leq b_i \leq \hat{b}_i + 1.96\ (\sigma_{\hat{b}1})$. When the level of significance $\alpha = 0.05$, then $Z\alpha/2 \approx 1.96$.

3.3.2 Estimation under ML system

i. **Uncertainty estimation in the regression model**

Under ML system, the study adopts the Quantile regression capability of LightGBM and RF regression models to estimate uncertainty. This allows the estimation of different quantiles of the target variable, which provides a measure of uncertainty. At each quantile interval prediction level, uncertainties are computed. A prediction interval (PI) at a specific quantile level represents the range within which the true target value is likely to lie with a chosen level of significance. The PI was used to construct the Prediction Interval Coverage Probability (PICP) will serves as a measure of uncertainty.

In order to explain the theoretical basis upon which uncertainty will be estimated in a ML modeling, assume a deterministic model of a real world M, making a prediction of output variable y* given input vector x(x ∈ X). Assume y to be the measurement made of an unknown true value with error e_y. Several kinds of errors propagate over the model M, in the course of predicting the observed output y. These errors could assume the following form:

$y= y^* + e_y = M(x, \theta) + e_s + e_\theta + e_x + e_y$ …………………….. (5)

Where:

θ = vector of model parameters

e_s = errors connected to the model structure M,

e_θ = error associated with model parameter θ

e_x = error associated with the input vector x

It is usually difficult to estimate the different errors separately without making certain assumptions, hence they are aggregated as one variable as given below:

$y = \hat{y} + e$ ………………………………….. (6)

Where:

$\hat{y}$ = the model output

e = total model error (residual)

Given the foregoing, the uncertainty estimation in this study will be the estimate of the uncertainty associated with the model structure M, and model parameter θ by analyzing model residuals e, which is a combined effect of all sources of error.

ii. **Uncertainty estimation in the classification model**

The standard deviation (SD) was adopted as a measure of uncertainty in the classification model. SD is used to determine the spread of the probability mass. As a measure of variability, a distribution's SD provides a valuable measure of uncertainty. To achieve this, the bootstrap sampling was applied in the classification model and the SD of each bootstrap sample ascertained. The values of the SD in relation to the accuracy level of the models formed a measure of uncertainty.

3.4 ML Modeling Flowchart

The process of applying ML in uncertainty quantification adopted in this study is depicted in figure 3 below. From a supervised ML scenario, the study splits into regression and classification cases. Data is cleaned and divided into training and test sets. Models are built and validated with the test dataset. Model evaluation, prediction and uncertainty estimation follow accordingly.

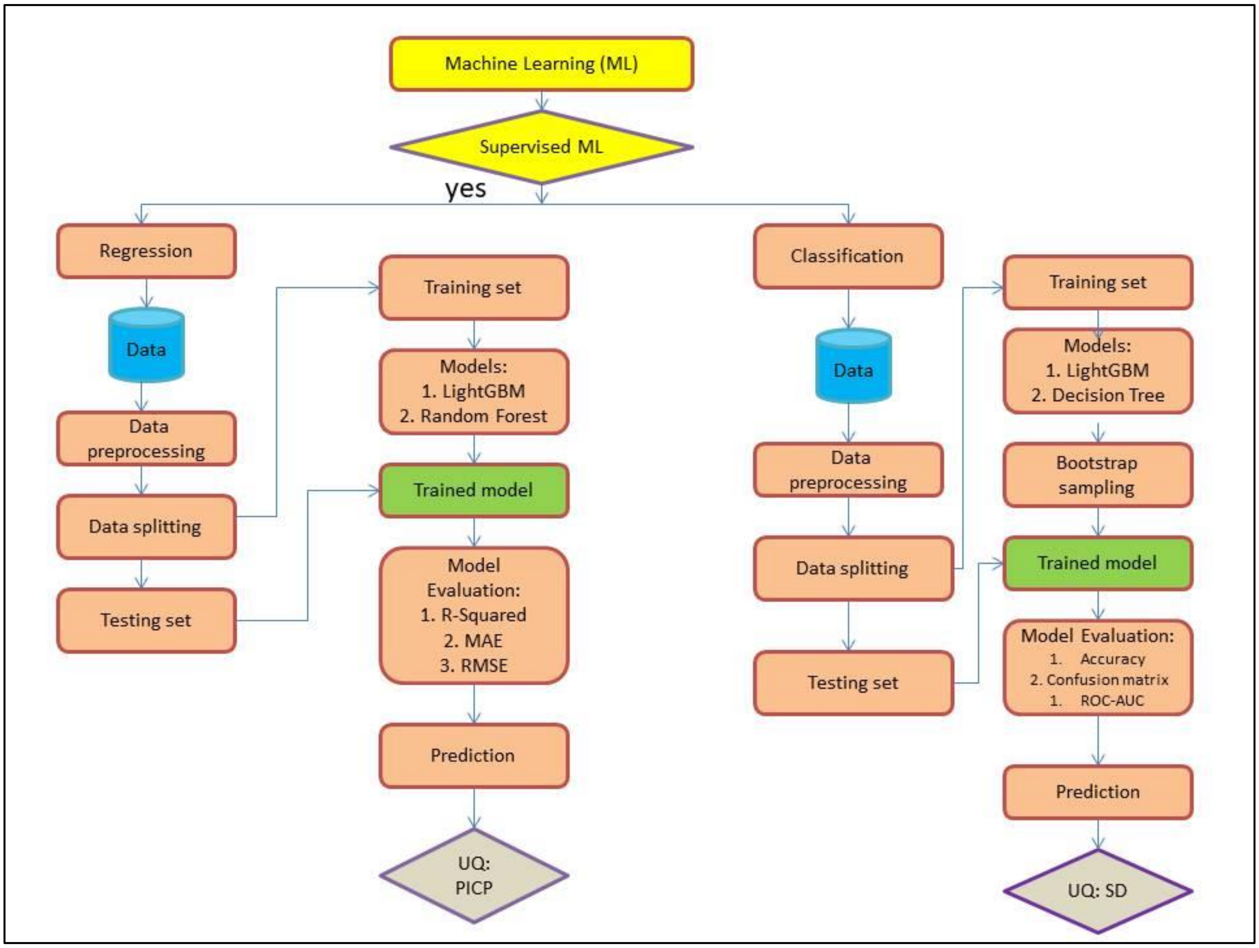


Figure 3: Machine learning Uncertainty quantification flowchart

3.5 Data, sources and implementation

This study employed secondary data and the analysis in this study were implemented in the Python programing module. Two data sets (dataset 1 and dataset 2) gotten from Kaggle website were used for analysis. Dataset 1 describes housing properties of Boston suburb and consists of 13 attributes (columns) and 506 rows. The problem here is to predict the house price (MEDV), given the features. The names of the features and the corresponding acronyms are presented in table 1 below:

Table 1: Dataset 1 Variables Description

S/N	COLUMN NAME	DESCRIPTION
1	CRIM	per capita crime rate by town
2	ZN	proportion of residential land zoned for lots over 25,000 sq.ft.
3	INDUS	proportion of non-retail business acres per town
4	CHAS	Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
5	NOX	nitric oxides concentration (parts per 10 million)
6	RM	average number of rooms per dwelling
7	AGE	proportion of owner-occupied units built prior to 1940
8	DIS	weighted distances to five Boston employment centres
9	RAD	index of accessibility to radial highways
10	TAX	full-value property-tax rate per 10,000usd
11	PTRATIO	pupil-teacher ratio by town
12	B	1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town
13	LSTAT	% lower status of the population
14	MEDV	Median value of owner-occupied homes in $1000's (house price)

Source: Author

Dataset 2 is about telecommunication customer behaviour and the purpose is to predict customer churn. It has insights into customer behaviour, and the development of focused customer retention programmes. It contains 21 columns and 7043 rows.

Each row represents a customer while each column contains customer’s attributes such as:

- Customers who left within the last month – the column is called Churn
- Services that each customer has signed up for – phone, multiple lines, internet, online security, online backup, device protection, tech support, and streaming TV and movies
- Customer account information – how long they’ve been a customer, contract, payment method, paperless billing, monthly charges, and total charges

- Demographic info about customers – gender, age range, and if they have partners and dependents

The interest in dataset 2 is to make a binary classification of the 'Churn' feature, predict the feature and estimate uncertainty of the prediction.

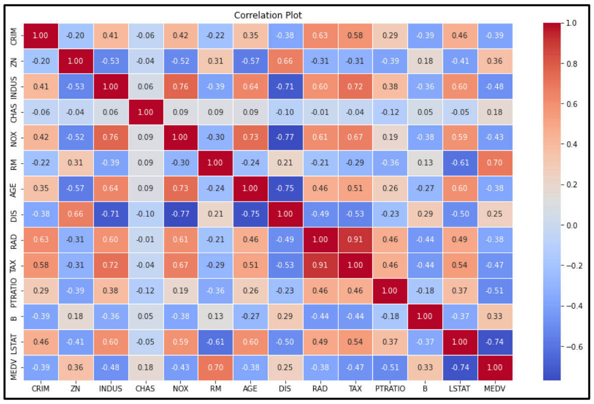
Exploratory Data Analysis (EDA) was performed on all the datasets for patterns, missing values and statistical properties/structures of the datasets. Data visualization was done to have a clearer way of to make it easier to identifying trends, patterns, and possibly to identify outliers. Dataset 1 was used for regression analysis while Dataset 2 was used for binary classification analysis.

4. Analysis of Results

4.1 Analysis of Regression Result

4.1.1 Regression EDA

The correlation matrix on dataset 1 was visualized to understand the strength and direction of relationships among the variables. This is shown in figure 4 with evidence that many of the variables have positive relationships. Similarly, a strong correlation was observed between RM and Price (MEDV) which was visualized in fig 5.



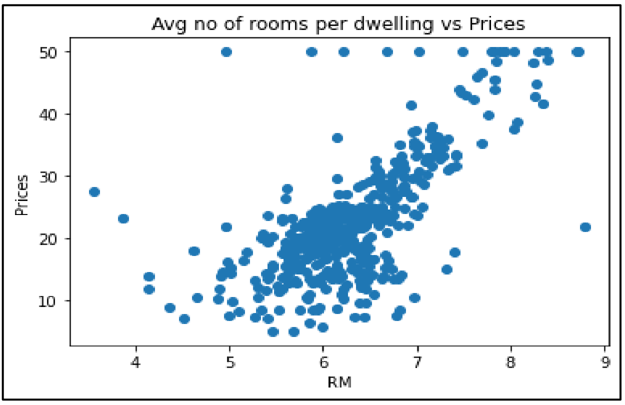


Figure 4: Correlation matrix of var in dataset 1.   Figure 5: Correlation between RM and Price

4.1.2 Summary of Regression results in the Frequentist method.

The summary of regression result using OLS is presented in table 2 below. The results show that the estimated parameters for INDUS with p-value of 0.738 and AGE with p-value of 0.958 are not statistically significant at the 5% level of significance. This is because their p-values are individually greater than 0.05. However, all the other variables are statistically significant at the 5% level given that their individual p-values are less than 0.05. The Adjusted R-square shows that about 73.4% of the variation in the response variable was explained by the model. Thus, the model has a moderately good fit. The F-statistic with a p-value of 6.72e-135 shows the overall model is statistically significant at the 5% level of significance.

Table 2: Summary of OLS result

OLS Regression result: Dep. Variable = MEDV								
					95% conf interval		**90% conf interval**	
S/N	**Variable Name**	**Estimated Coefficient**	**Std Error**	**P-value**	**lower**	**upper**	**lower**	**upper**
1	CRIM	-0.108	0.033	0.001	-0.173	-0.043	-0.162	-0.054
2	ZN	0.046	0.014	0.001	0.019	0.073	0.024	0.069
3	INDUS	**0.021**	**0.061**	0.738	-0.100	0.141	-0.081	0.122
4	CHAS	2.687	0.862	0.002	0.994	4.380	1.267	4.107
5	NOX	-17.767	3.820	0.000	-25.272	-10.262	-24.061	-11.472
6	RM	3.810	0.418	0.000	2.989	4.631	3.121	4.499
7	AGE	0.001	0.013	0.958	-0.025	0.027	-0.021	0.022
8	DIS	-1.476	0.199	0.000	-1.867	-1.084	-1.804	-1.147
9	RAD	0.306	0.066	0.000	0.176	0.436	0.197	0.415
10	TAX	-0.012	0.004	0.001	-0.020	-0.005	-0.019	-0.006
11	PTRATIO	-0.953	0.131	0.000	-1.210	-0.696	-1.168	-0.737
12	B	0.009	0.003	0.001	0.004	0.015	0.005	0.014
13	LSTAT	-0.525	0.051	0.000	-0.624	-0.425	-0.608	-0.441
R-Squared=0.741								
Adj R-squared=0.734; MAE= 3.27; MSE=21.89; RMSE=4.68								
F-Stat=108.1								
prob(F-stat)= 6.72e-135								

Source: Author's computation

4.1.3 Diagnostics of OLS model residuals

Diagnostic analysis was performed to investigate the properties of the residuals from the estimated model so as to make conclusions about the violations or otherwise of the assumptions underlying the multiple regression analysis. The Variance Inflation Factor (VIF) was used to test for Multicollinearity and the result is presented in table 3 below. Similarly, the Residuals were plotted against theoretical values and the results are presented below in Figure. 6.

Table 3: VIF result

Variable	**CRIM**	**ZN**	**INDUS**	**CHAS**	**NOX**	**RM**	**AGE**	**DIS**	**RAD**	**TAX**	**PTRATIO**	**B**	**LSTAT**
VIF	1.79	2.30	3.99	1.07	4.39	1.93	3.10	3.96	7.48	9.01	1.80	1.35	2.94

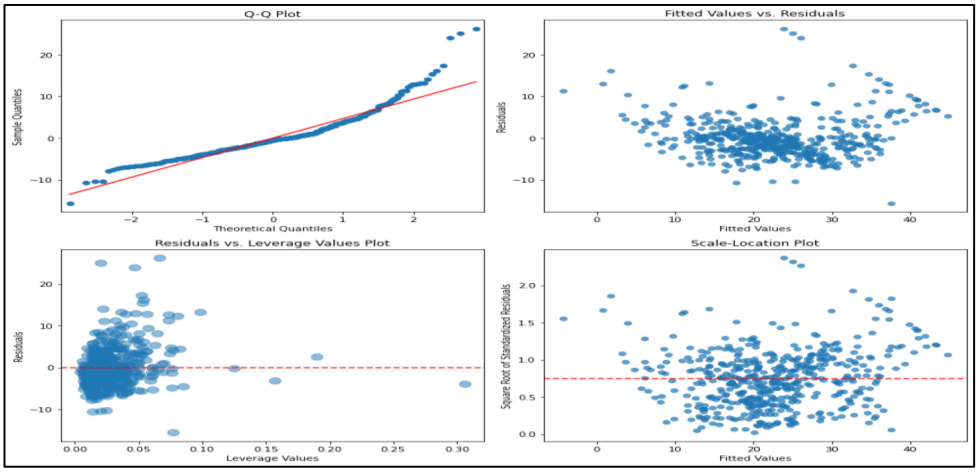


Figure 6: Residuals diagnostics results

The result from the multicollinearity test (Table 3) shows the absence of serious multicolinearity among the explanatory variables. Apart from RAD and TAX with VIF scores of 7.48 and 9.01 respectively, all the other variables have scores less than 5.0, indicating no multicolinearity. However, multicolinearity is still at tolerable levels for RAD and TAX since the values are each less than 10.

Diagnostic analysis performed to investigate the properties of the residuals from the estimated showed none violation of the assumptions underlying the regression model. Thus, the Residuals were plotted against theoretical values as shown in shown in Figure 6. Examination of these plots shows that the assumptions of linearity, normality of the error term and homoscedasticity were satisfied.

4.1.4 Uncertainty estimation in the traditional model

Table 2 shows the lower and upper boundary values for the parameter estimates at both the 95% and 90% confidence interval which is applied to measure uncertainty. For instance, given the estimated parameter of the variable CRIM, the results show we are 95% confident that the true population parameter lies within the range of [-0.173, -0.043]. Similarly given the estimated parameter of the variable ZN, we are 90% confident that the true population parameter lies within the range [0.024, 0.069]. The same analogy goes for all the other estimated parameters in the model.

4.1.5 Model Performance indicators for the ML models

Using training - test dataset ratio of 70%:30%, table 4 shows hyperparameter tuning values for the two ML regression models. The tuned hyperparameters for LightGBM and FR have the same tree-based tuning parameters, learning rate and number of estimators. However, while maximum depth for LightGBM is 5, there is no limit for RF.

Table 4: Hyperparameter tuning

Method	no of leaves	learning rate	no of estimators	maximum depth	Random state
LightGBM	31	0.1	20	5	42
Random Forest	31	0.1	20	-	42

Source: Author's computation

Table 5: ML regression models' evaluation indicators

ML model diagnostics					
		Model Evaluation Metrics			
S/N	Model	R-Squared	MAE	MSE	RMSE
1	Lightgbm	0.828	2.069	12.825	3.581
2	Random Forest	0.862	2.122	10.284	3.207

Source: Author's computation

Similarly table 5 shows the performance indicators for the models. The R-Squared of 0.828 for LightGBM and 0.862 for RF show that 82.8% and 86.2% variation in the response variable was explained by the models respectively. This shows a good fit for both models, however RF outperformed LightGBM. The Mean Absolute Error (MAE) predictive ability indicator shows that LightGBM with a value of 2.069 has a better predictive ability than RF with a value of 2.122. However, in terms of the other metrics (MSE and RMSE), RF is a better model.

4.1.6 Uncertainty evaluation indicators: regression model

Table 6, figure 7 and figure 8 present indicators of uncertainty in the ML regression models. Quantile regression was used to estimate the lower and upper quantiles from which the Prediction Intervals (PIs) were constructed. The Prediction Interval Coverage probability (PICP) and the Mean Prediction Interval Width (MPIW) were estimated from the test dataset. PICP was measured by calculating the number of observations covered by the constructed PIs while the MPIW estimates the average width of PIs. MPIW provides insights into how wide or narrow the prediction intervals are on average. Uncertainty in the model was assesses with both the PICP and MPIW respectively. A well calibrated PI should be as narrow as possible while capturing a specified portion of data, say 90% or 95% of the data. So, when it is asked to output a 90% PI, an accurate uncertainty estimation method should produce a small MPIW value while ensuring its PICP around 90%.

Table 6: 90% PICP and MPIW

Method	PICP	MPIW
LightGBM	0.875	6.739
Random forest	0.914	9.408

Source: Author's computation

From the result, Random forest has a PICP of 0.914 while the LightGBM 0.875 at the 90% prediction level. This means that 91.4% and 87.5% of the observations fall within the 90% prediction level for Random Forest and LightGBM respectively. Similarly, LightGBM has an MPIW of 6.739 which is lower than Random forest's MPIW of 9.408. This is clearly visualized in figure 7 and figure 8. The yellow shaded area (prediction interval) in Fig 8 representing Random forest is wider than in fig 7 representing Light prediction.

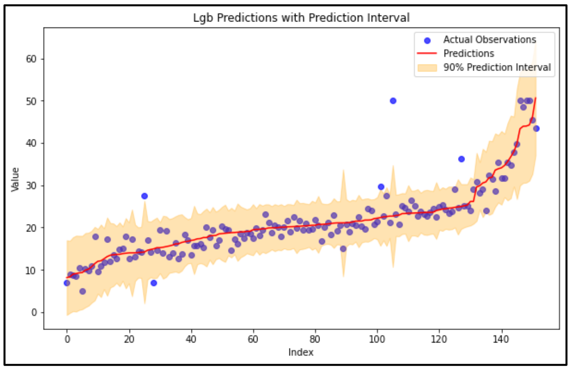


Figure 7: LigbtGBM prediction

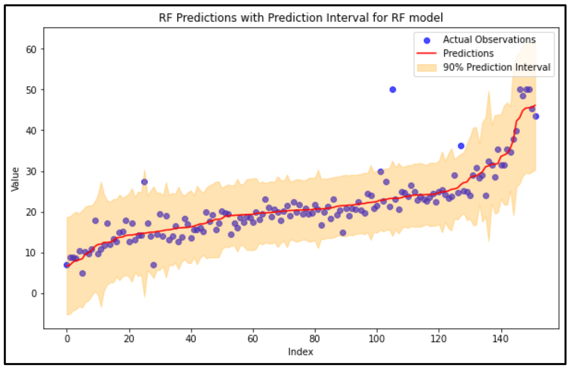


Figure 8: Random Forest prediction

4.2 Analysis of Classification Result

4.2.1 Classification model evaluation result

The results of the binary classification using Decision Tree (DT) classifier and LightGBM classifiers are presented in table 7, Figure 9, Figure 10, Figure 11 and Figure 12 respectively. The table and graphs show the confusion matrix and the ROC (Receiver Operating Characteristics) –Area under the curve (AUC).

Table 7: Classification metrics and ROC-AUC

Model evaluation metrics					
	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Decision Tree	0.79	0.65	0.49	0.56	0.83
LightGBM	0.81	0.68	0.54	0.60	0.86

Source: Author's computation

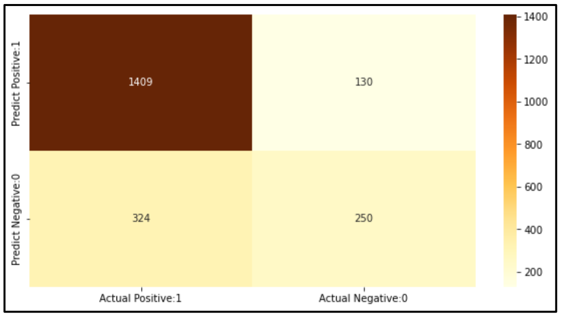


Figure 9: DT Confusion matrix

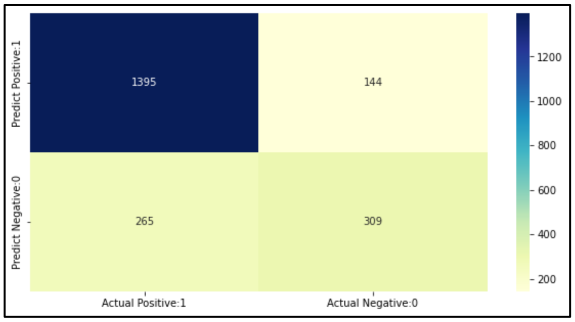


Figure .10 LGBM Confusion matrix

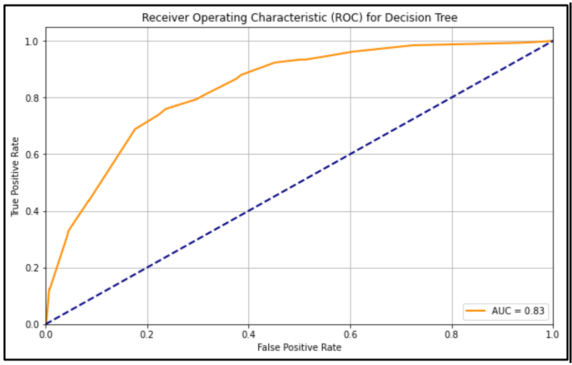


Figure11: Decision Tree ROC-AUC

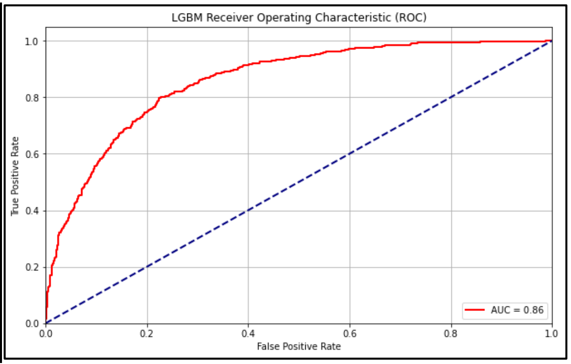


Figure 12: LGBM ROC-AUC

LightGBM has an accuracy level of 81% while DT accuracy is 79%. Similarly, LightGBM has an ROC-AUC of 86% which is higher than DT's ROC-AUC of 81%. The Precision and Recall levels for LightGBM are 68% and 65% respectively. These figures are higher than the corresponding figures for DT.

4.2.2 Uncertainty evaluation: classification model

Uncertainty in the classification model was evaluated with the standard deviation distributions of the Bootstrap sample models for both the LightGBM and DT. These are depicted in Fig 13 and Figure 14. Similarly, the probability density plots for both models were also evaluated as shown in figures 15 and 16 respectively for class label 1 (the existence of customer churn).

This analysis considered a case of class purity. Only class label 1 was therefore analysed and a discreteized accuracy/SD plot was generated. Given that each leaf contains only one class, when a test sample falls to a particular leaf it assumes the same label with a 100% probability. Since this is a binary situation the estimation of mean and SD will result in values which display little variation. In figure 13 and figure 14, different colour palettes were used to show various SD at various accuracy levels for both

LightGBM and Decision Tree Models. Along the curve, the SD/Accuracy combination varies. For instance, at accuracy level 100%, the SD=0, while at accuracy level 80%, SD= 0.4. The average accuracy for LightGBM is 79% while it is 72% for Decision Tree. The relationship between SD and accuracy is very useful even when the classifiers' overall accuracy is reduced. SD of prediction probability distributions correlates with accuracy

Avg Accuracy: 72%; Avg SD= 0.32

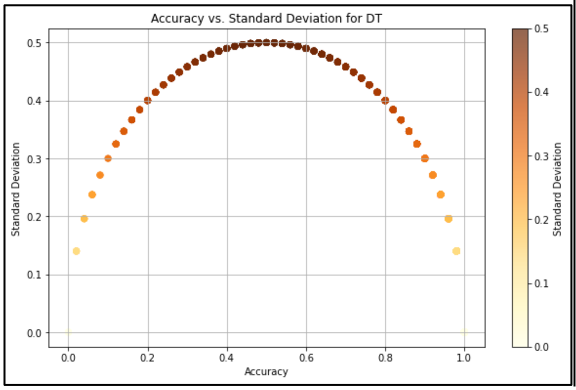


Figure 13: SD v Accuracy for DT model

Avg Accuracy: 79%; Avg SD= 0.15

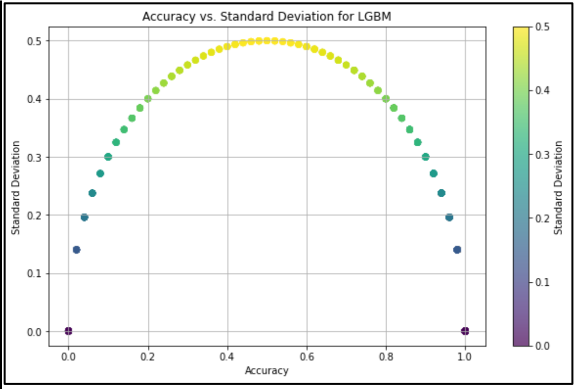
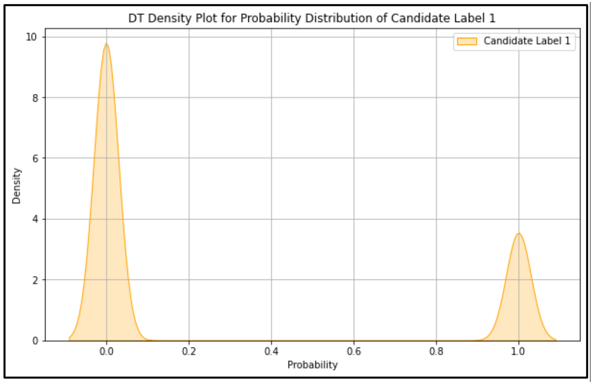


Figure 14: SD v Accuracy for LGBT model



Figuure15: Prob density plot for DT model

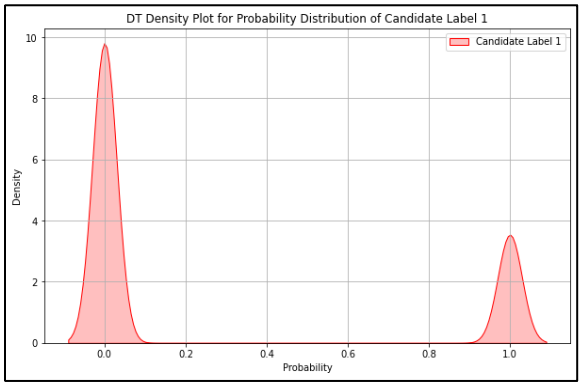


Figure 16: Prob density plot for LGBM model

5. Discussion of Results

5.1 Discussion of model Results

The 74.1% predictive ability of the traditional model evidenced by the R-Squared indictes the model has a good fit. However, this could be improved by enhancing the data quality- increasing the data points, removing anomalies such as outliers, removing extraneous variables or adding new features. This could boost the performance of the model. Furthermore, a typical method for quantifying uncertainty is the calculation of confidence intervals as enunciated by the Frequentists. The results from the traditional model are consistent with the classical Frequentist model. The standard errors, p-value and confidence intervals constructed show the uncertainty in the estimation of the model parameters in line with the theoretical underpinnings of the model. Under the assumption of normality, the confidence interval as shown in table 2, contains the true posterior probability with probability (95% or 90%) = 1- α. The result is also consistent with studies such as Darling, M. C, (2019), Kerns, (2017) and Rostron *et al.*, (2020).

In the ML regression models, both Random Forest and LightGBM showed high predictive powers with high R-Squared of 0.828 and 0.862 respectively. Other model evaluation metrics such as MAE, MSE and RMSE also indicated the models have reliable predictive abilities. The introduction of Quantile regression which led to estimation of prediction intervals provided a footing for the estimation of PICP and MPIW, thus providing a route for the estimation of uncertainties in the models. Results showed uncertainties are estimated within the PICP bands with Random forest performing better with larger coverage. Results emanating from these models are consistent with a-priori expectation and consistent with studies such as Xiaozhe *et al.*, (2023); and Rahmati, et. al (2019). Similar hyperparameter tuning for the two models provided a good basis for comparing the outcome of the two models. Consequently, evidence shows Random Forest with PICP of 0.914 performed better than LightGBM with PICP of 0.875 (table 6).

In the ML classification models, results show LightGBM performed better than DT models both in accuracy of predicting customer churn and AUC. Furthermore, 100 bootstrap samples were trained under LightGBM and DT models. When these models make predictions on the test data, the variance among their predictions were calculated and this reflects the spread or diversity of the individual model predictions. The variability in the prediction of these samples was estimated using the SD as shown in Figure 13 and figure 14. This shows how confident the ensemble is in its prediction and thus provided a measure of the overall uncertainty.

Figures 13 and 14 show that in this binary case, the maximum SD (0.5) occurs when the model's predicted probability is close to 0.5, indicating that the model is less certain about the class assignment. As the predicted probabilities approach 0 or 1, the SD tends to decrease, reflecting higher confidence in the prediction. The result shows lower variability in the prediction of LightGBM model with average SD of 0.15 (Fig 14). In other words, uncertainty is lower in LightGBM model than when making prediction with DT model. This result is consistent with Darling, M. C, (2019); Darling and Stracuzzi, (2018); and Gal, Y. and Ghahramani, Z. (2016).

5.2 Traditional uncertainty estimation vs. ML estimation results

The traditional statistical method makes use of confidence interval in the assessment of uncertainty. In doing this, it makes a-prior assumptions about the error term such as assuming normality. Estimates made under this are not point estimate but a range of values within which the population values will lie. Alternatively, in ML, while PI gives interval estimates, the coverage probability (PICP) gives a percentage assessment of the prediction interval coverage, aimed at assessing the validity of prediction intervals. It assesses how well the prediction intervals behave in practice. The MIPW is used to evaluate the optimality. For instance, in the result for the traditional model, the true population parameter for the variable CRIM, lies within the range of [-0.162, -0.054] at 1- α =0.90 while under ML model, the PICP shows that for LightGBM 87.5% of all the observations are covered by the PI. Similarly, under ML all the point estimate evaluation metrics such as MAE, MSE and RMSE are lower than under the traditional method. This shows that ML method is better than the traditional Frequestist method as the former brings some sort of precision and reliability in uncertainty estimation.

6. Conclusion and Recommendation

6.1 Conclusion

The aim of this study was to investigate how uncertainty is quantified in Machine Learning models. Against this backdrop, uncertainty quantification in the traditional statistical system was examined using the Frequentist method and the OLS regression model. Similarly, supervised ML modeling was set up using the regression and the classification models. Two ML algorithms were adopted in each case: LightGBM and Random forest for the regression model; DT and LightGBM for the classification model respectively. The dataset for the ML models was split in the training/test ratio of 70:30. The response variables were house price in the regression models and customer churn in the classification models.

The use of confidence intervals to measure uncertainty in the traditional system was examined. Confidence intervals were constructed at the 90% and 95% levels respectively. The OLS results showed good predictive capabilities as the model explained 74.1% variations in the response variable. In addition, model diagnostics results showed that the assumptions about the model residuals were satisfied. It was however observed that data quality could have impacted the explanatory ability of the model, giving it a moderately high R-Squared.

Results from the ML regression models indicated high explanatory abilities for the RF and LightGBM regressors, as well as high predictive abilities given their R-squared, MSE, MAE and RMSE respectively. However, RF produced better point estimation results with R-squared of 86.2% and RMSE = 3.207. Similarly, RF with higher PICP (0.914) and MPIW (9.409) values provided a better way of accounting for uncertainties in the prediction than LightGBM.

The results of the ML regression models were generally better than those produced by the traditional regression models. The ML models produced higher explanatory ability evidenced by higher R-squared and lower MAE/MSE. Uncertainty estimation under ML was also preferred as the PICP is considered by this study a better indicator based on the available results.

The performance of the binary classification models followed similar fashion as the regression models. LightGBM produced an accuracy level of 81% with ROC-AUC of 86% which were better than the output of DT. In assessing uncertainties, bootstrapping was adopted with 100 boothstrap samples. The hyperparameters of the models were tuned such that the trees were allowed to grow until the leaves achieved class purity. SD from the bootstrap samples provided a measure of uncertainty in the models' predictions. The relationship between SD and accuracy indicates that SD of prediction probability distributions correlates with accuracy. LightGBM model performed better as it showed less variability in its prediction evidenced by lower average SD of 0.15. Besides, in a classification problem, estimating a model's uncertainty in the prediction of the label of a sample allows us to infer a degree of confidence.

In general, the empirical analysis showed that machine learning modeling performed better in investigating uncertainty. ML models provided interval uncertainty estimation and PICP which show how many observations fall into the prediction space. Besides ML models account for both the Epistemic and Aleatoral uncertainties. The definition and calculation of uncertainty depends on both the context and the modeling algorithm. Based on the empirical outcome of this study, PICP, MIPW and SD have been found very useful in quantifying uncertainty in ML model.

6.2 Recommendation

Based on the outcome of this study, the following recommendations are offered.

i. ML model developers should provide uncertainty estimates accompanying each model prediction at all times. This will help to deepen users' trust in ML model outputs as well as support reliable decision making.

ii. Using uncertainty estimation to increase trust requires articulating appropriate methods, diagnostics and visualization pathways. Therefore, in deciding which UQ method to adopt several competing methods should be evaluated so as to choose the best in view of the peculiarity of the scenario under consideration and data quality.

iii. ML models should take preference over traditional methods in both predictions and uncertainty estimation. Business should adopt ML models over the traditional methods.

iv. Random forest algorithm and LightGBM should be adopted in uncertainty modeling situations as they provide better results.

6.3 Future Work

Quantifying a model's uncertainty elicits a measure of confidence in the model's predictions. Uncertainty estimates is necessary in decision making. It could also be used to determine if alternate models or different data/hyperparameters are required. Based on the foregoing, future work should consider the following areas:

i. **ML methods:** This study adopted LightGBM, RF and DT ML algorithms. Other ML models should be examined to see how their results compare with this study.
ii. **Multiple classes:** In the classification problem, this study adopted a binary case. It will be necessary to consider multi-class classification in future works to see how results from multiple classes could compare with this study.
iii. **Hyperparameter tuning:** The impact of hyperparameter tuning in uncertainty quantification is a good area of interest. This could potentially produce differing results with consequential implications for model predictions.
iv. **Uncertainty quantification method:** This study adopted the ensemble method, however future studies should cover the Monte Carlo dropout and the Gaussian processes. The implication of differing outcomes will produce a strong basis for model selection and explainability.

References

Abdar, M., Pourpanah, F., Hussain, S., Rezazadegan, D., Liu, L., Ghavamzadeh, M., Fieguth, P., Cao, X., Khosravi, A., Acharya, U.R., Makarenkov, V. and Nahavandi, S. (2021). A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information Fusion*,76, pp.243–297. doi:https://doi.org/10.1016/j.inffus.2021.05.008.

Akaike, H. (1973). *Information Theory and an Extension of the Maximum Likelihood Principle. In B. N. Petrov, & F. Csaki (Eds.), Proceedings of the 2nd International Symposium on Information Theory (pp. 267-281). Budapest Akademiai Kiado.* -Available at: https://www.scirp.org/(S(vtj3fa45qm1ean45vvffcz55))/reference/ReferencesPapers.aspx?ReferenceID=1209327.

Antorán, J., Bhatt, U., Adel, T., Weller, A. and Miguel Hernández-Lobato, J. (n.d.). *Published as a conference paper at ICLR 2021 Getting A Clue: A Method For Explaining Uncertainty Estimates*. Available at: https://openreview.net/pdf?id=XSLF1XFq5h [Accessed 18 Aug. 2023].

Arkov, V. (2022). *Uncertainty estimation in mechanical and electrical engineering, in Proc. 2021 International Conference on Electrotechnical Complexes and Systems ICOECS, 436–440. arXiv.org. doi:https://doi.org/10.48550/arXiv.2206.01749.*

Bayes, T., (1763). *An essay towards solving a problem in the doctrine of chances*. Philosophical Transactions of the Royal Society of London 53, pp. 370–418.
Butvinik D, (2022). Uncertainty Quantification in Artificial Intelligence-based Systems.KD Nugetts. [Accessed 18 Aug. 2023]. https://www.kdnuggets.com/2022/04/uncertainty-quantification-artificial-intelligencebased-systems.html

Caldeira, J. and Nord, B. (2020). Deeply uncertain: comparing methods of uncertainty quantification in deep learning algorithms. *Machine Learning: Science and Technology*, 2(1), p.015002. doi:https://doi.org/10.1088/2632-2153/aba6f3.

Chen, T., Dey, B., Ghosh, A., Kagan, M., Nord, B. and Ramachandra, N. (2022). *Interpretable Uncertainty Quantification in AI for HEP*. Submitted to the Proceedings of the US Community Study on the Future of Particle Physics (Snowmass 2021). Available at: https://www.slac.stanford.edu/econf/C210711/papers/2208.03284.pdf [Accessed 18 Aug. 2023].

Choubineh, A., Chen, J., Coenen, F. and Ma, F. (2023). Applying Monte Carlo Dropout to Quantify the Uncertainty of Skip Connection-Based Convolutional Neural Networks Optimized by Big Data. *Electronics*, 12(6), p.1453. doi:https://doi.org/10.3390/electronics12061453.

Clyde, M. and George, E.I. (2004). Model Uncertainty. *Statistical Science*, [online] 19(1), pp.81–94. Available at: https://www.jstor.org/stable/4144374 [Accessed 18 Aug. 2023].
Darling, M. (2019). Using Uncertainty To Interpret Supervised Machine Learning Predictions. *Electrical and Computer Engineering ETDs*. [online] Available at: https://digitalrepository.unm.edu/ece_etds/485 [Accessed 28 Aug. 2023].

Dietterich, T.G. (2000). Ensemble Methods in Machine Learning. In: Multiple Classifier Systems. MCS 2000. Lecture Notes in Computer Science, 1857. Springer, Berlin, Heidelberg. https://doi.org/10.1007/3-540-45014-9_1

Edanz, (2023). Bayesian Analysis in Systematic Reviews and Meta-analyses. Accessed 12/07/2023. https://learning.edanz.com/frequentist-bayesian-statistics/.[Accessed 28 Aug. 2023]
Efron, B. (1979). Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics*, 7(1), pp.1–26. doi:https://doi.org/10.1214/aos/1176344552.

Frau, L., Susto, G.A., Barbariol, T. and Feltresi, E. (2021). Uncertainty Estimation for Machine Learning Models in Multiphase Flow Applications. *Informatics*, 8(3), p.58. doi:https://doi.org/10.3390/informatics8030058.

Freni, G. and Mannina, G. (2010). Bayesian approach for uncertainty quantification in water quality modelling: The influence of prior distribution. 392(1-2), pp.31–39. doi:https://doi.org/10.1016/j.jhydrol.2010.07.043.

Gal, Y. and Ghahramani, Z. (2016). *Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning*. [online] proceedings.mlr.press. Available at: https://proceedings.mlr.press/v48/gal16.html [Accessed 28 Aug. 2023].

Giaquinto, N., Fabbiano, L. , Trotta, A, and Vacca, G (2014)About the Frequentist and the Bayesian Approach to Uncertainty: 20th IMEKO TC4 International Symposium and 18th International Workshop on ADC Modelling and Testing Research on Electric and Electronic Measurement for the Economic Upturn Benevento, Italy, September 15-17, 2014

Hariri, R.H., Fredericks, E.M. and Bowers, K.M. (2019). Uncertainty in Big Data analytics: survey, opportunities, and Challenges. *Journal of Big Data*, 6(1), pp.1–16. doi:https://doi.org/10.1186/s40537-019-0206-3.

Hoffmann, L., Fortmeier, I. and Elster, C.(2021) Uncertainty quantification by ensemble learning for computational optical form measurements. Mach. Learn.: Sci. Technol. 2 (2021) 035030
https://doi.org/10.1007/s10994-021-05946-3

Hullermeier, E. and Waegeman, W. (2021). Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods. *Machine Learning*, [online] 110(3), pp.457–506. doi:https://doi.org/10.1007/s10994-021-05946-3.

Indrayan, A. (2020) Aleatory and epistemic uncertainties can completely derail medical research results Journal of Postgradute Medicine. 66(2): 94–98. doi: 10.4103/jpgm.JPGM_585_19

Ivanov, G.(2020).Uncertainty Assessment of Predictions with Bayesian Inference An introduction to computational Bayesian statistics. https://towardsdatascience.com/uncertainty-quantification-of-predictions-with-bayesian-inference-6192e31a9fa9#:~:text=Bayesian%20statistics%20is%20a%20powerful,anomalous%20samples%20in%20test%20data.

Kazmierczak, N. P., Joyce A. Chew, J. A and Griend, D. A. (2022).Bootstrap methods for quantifying the uncertainty of binding constants in the hard modeling of spectrophotometric titration data. Analytica Chimica Acta -1227

Kendall, A., & Gal, Y. (2017). What uncertainties do we need in Bayesian deep learning for computer vision? In *Proceedings of NIPS, advances in neural information processing systems*, pp. 5574–5584

Kerns, L. (2016). Confidence bands for the logistic and probit regression models over intervals. *Communications in Statistics - Theory and Methods*, 46(8), pp.3878–3890. doi:https://doi.org/10.1080/03610926.2015.1073319.

Kiureghian, A.D. and Ditlevsen, O. (2009). Aleatory or epistemic? Does it matter? *Structural Safety*, [online] 31(2), p.105. Available at: https://www.academia.edu/98271679/Aleatory_or_epistemic_Does_it_matter [Accessed 11 Aug. 2023].

Konig, H.M.T., Hoos, H.H.and Rijn, J.N. van.(2020).Towards algorithm-agnostic uncertainty estimation: predicting classification error in an automated machine learning setting. 7th ICML Workshop on Automated Machine Learning (AutoML 2020). Leiden University

Lane, D. and Maxfield, R. (2004). Ontological uncertainty and innovation. *Journal of Evolutionary Economics*, [online] 15(1), pp.3–50. Available at: https://ideas.repec.org/a/spr/joevec/v15y2004i1p3-50.html [Accessed 24 Aug. 2023].

MacAskill, W., Bykvist, K. and Ord, T.(2020) *Moral Uncertainty* Oxford University Press, Oxford.
Nagl, M., Nagl, M. and Rösch, D. (2022). Quantifying uncertainty of machine learning methods for loss given default. *Frontiers in Applied Mathematics and Statistics*, 8. doi:https://doi.org/10.3389/fams.2022.1076083.

National Research Council. (2012). *Assessing the Reliability of Complex Models: Mathematical and Statistical Foundations of Verification, Validation, and Uncertainty Quantification*. Washington, DC: The National Academies Press. https://doi.org/10.17226/13395.

Pevec, D and Kononenko, I. (2015) Prediction intervals in supervised learning for model evaluation and discrimination. eng. In: *Applied Intelligence* 42.4 (2015), pp. 790–804.
Rahman, M (2023) Monte Carlo Dropout for Uncertainty Estimation in Deep Learning. Accessed 12/07/2023: https://rmoklesur.medium.com/monte-carlo-dropout-for-uncertainty-estimation-in-deep-learning-model-e1b4b9254d4e

Rahmati, O., Choubin, B., Fathabadi, A., Coulon, F., Soltani, E., Shahabi, H., Mollaefar, E., Tiefenbacher, J., Cipullo, S., Ahmad, B. B., & Tien Bui, D. (2019). Predicting uncertainty of machine learning models for modelling nitrate pollution of groundwater using quantile regression and UNEEC methods. *Science of the Total Environment*, *688*, pp.855–866. https://doi.org/10.1016/j.scitotenv.2019.06.320

Rasmussen, C. E. and Williams, C. K. (2006) Gaussian Processes for Machine Learning, the MIT Press, 2006, ISBN 026218253X. c 2006 Massachusetts Institute of Technology. Available at: https://gaussianprocess.org/gpml/chapters/RW.pdf.

Rasmussen, C.E.(2003) Gaussian processes in machine learning. In Summer School on Machine Learning; Springer: Berlin, Heidelberg, 2003; pp. 63–71.

Rostron, P.D., Fearn, T. and Ramsey, M.H. (2020). Confidence intervals for robust estimates of measurement uncertainty. *Accreditation and Quality Assurance*, 25(2), pp.107–119. doi:https://doi.org/10.1007/s00769-019-01417-4.

Schmitt, L. (2013). *Ontological Uncertainty and the Expansion of Experience*. [online] Inovo. Available at: https://theinovogroup.com/ontological-uncertainty-and-the-expansion-of-experience/ [Accessed 28 Aug. 2023].

Senge *et al.* (2014).Reliable classification: Learning classifiers that distinguish aleatoric and epistemic uncertainty.*Information Sciences*, *255,* pp.16–29.

Solomatine, D.P. and Shrestha, D.L. (2009). A novel method to estimate model uncertainty using machine learning techniques. *Water Resources Research*, 45(12). doi:https://doi.org/10.1029/2008wr006839.

Souza,R.R., Dorn, A., Piringer, B. and E. Wandl-Vogt, E. (2019.Towards A Taxonomy of Uncertainties: Analysing Sources of Spatio-Temporal Uncertainty on the Example of Non-Standard German Corpora. *Informatics*, 6(3), p.34. doi:https://doi.org/10.3390/informatics6030034.

Tavazza, F., DeCost, B. and Choudhary, K. (2021). Uncertainty Prediction for Machine Learning Models of Material Properties. *ACS Omega*, 6(48), pp.32431–32440. doi:https://doi.org/10.1021/acsomega.1c03752.

Thompson, A., Jagan, K., Sundar, A., Khatry, R., Donlevy, J., Thomas, S. and Harris, P. (2021). *NPL REPORT MS 34 UNCERTAINTY EVALUATION FOR MACHINE LEARNING*. [online] Available at: https://eprintspublications.npl.co.uk/9306/1/MS34.pdf [Accessed 18 Aug. 2023].

Volodina, V. and Challenor, P. (2021). The importance of uncertainty quantification in model reproducibility. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 379(2197). doi:https://doi.org/10.1098/rsta.2020.0071.

William, M., Krister, B., and Ord, T. (2020). *Moral Uncertainty*. Great Clarendon Street, Oxford, OX2 6DP, 1 United Kingdom

Willink, R and White, R (2012). Disentangling Classical and Bayesian Approaches to Uncertainty. Accessed on 20/07/2023 https://www.scribd.com/document/336247071/Disentangling-uncertainty-v14-pdf#

Yang, J., Reichert, P., Abbaspour, K.C., Xia, J., Yang, H., (2008). Comparing uncertainty analysis techniques for a SWAT application to the Chaohe Basin in China. *Journal of Hydrology 358 (1–2), pp. 1–23.*

Yin, X., Fallah-Shorshani, M., Mcconnell, R., Fruin, S., Chiang, Y.-Y. and Franklin, M. (2023). *Quantile Extreme Gradient Boosting For Uncertainty Quantification*. [online] Available at: https://arxiv.org/pdf/2304.11732.pdf.