Visual Tools for Understanding Regression Black-Box Models

Maria Inês Costa Areosa Rodrigues

Thesis to obtain the Master of Science Degree in

Aerospace Engineering

Supervisor(s):	Prof Luís Fernando Rainho Alves Torgo
	Prof Luís Manuel Marques Custódio

Examination Committee

Chairperson: Prof José Fernando Alves da Silva
Supervisor: Luís Fernando Rainho Alves Torgo
Member of the Committee: Mário Alexandre Teles de Figueiredo

November 2019

We lean forward to the next crazy venture beneath the skies

Jack Kerouac, 1957

Acknowledgments

I would like to express my deep gratitude to my supervisors, Luís Torgo and Luís Custódio for the guidance, motivation and support throughout the entire project. I am incredibly thankful for all the knowledge imparted. Moreover, I would like to thank Dr. Kristina Boerder for the help with the motivating case study.

My sincere thanks to all the extraordinary people I have befriended along the way. I am genuinely grateful for all the kindness, encouragement and wisdom. To all my friends and family, thank you for making anywhere feel like home.

Finally, I wish to express my profound gratitude to my parents for always teaching me in such a fantastic and sometimes unusual way. Thank your for all the opportunities and for the unconditional support. Every accomplishment of mine, will always be yours.

This work was partially funded by FCT - Fundação para Ciência e Tecnologia (Portuguese Science and Technology Foundation) through the NITROLIMIT project (PTDC/CTA-AMB/30997/2017).

Resumo

A falta de transparência é, atualmente, uma das principais barreiras à adoção de técnicas de aprendizagem automática, apesar do excecional desempenho de algoritmos mais recentes. Aquando da tomada de decisões importantes e dispendiosas, os utilizadores apenas conseguem depositar total confiança nas predições de um modelo se entrevirem o seu funcionamento. Por conseguinte, explicar modelos opacos (caixas pretas) tem-se tornado num dos tópicos em voga na investigação em aprendizagem automática. Existem variados métodos que podem ser implementados para perceber um modelo. Este trabalho enquadra-se na explicação de modelos preditivos de regressão através de ferramentas visuais, visto que estas são mais adequadas para transmitir informação a utilizadores com reduzido conhecimento técnico.

Apesar da maior parte dos métodos existentes analisar apenas a saída dos modelos, defendemos que avaliar o desempenho de um modelo é também de extrema importância. Por esta razão, as contribuições deste trabalho inserem-se neste último aspeto da explicabilidade. Neste trabalho desenvolvemos uma nova perspetiva, que inspeciona os riscos inerentes a usar um modelo de regressão opaco num certo domínio de preditores. Assim, é proposto e avaliado um conjunto de ferramentas que transmitem visualmente a relação entre o erro esperado e os valores de um preditor.

Por fim, é abordado um problema concreto, em que utilizamos ferramentas de explicabilidade para compreender que fatores influenciam o esforço de pesca em áreas marinhas protegidas de grande escala. Nesta análise são também comparados alguns modelos preditivos para selecionar o mais adequado ao problema e posteriormente é feita uma análise geral ao desempenho do modelo escolhido.

Palavras-chave: Explicabilidade, Caixa preta, Regressão, Desempenho, Responsabilidade, Áreas marinhas protegidas de grande escala

Abstract

Lack of transparency has become a significant barrier to the widespread adoption of Machine Learning techniques in many areas of human society, despite the outstanding performance of recent algorithms in terms of accuracy. When accounting for important and costly decisions, end users need to understand the model to be able to rely on the predictions. In that regard, explaining black-box models has become a hot topic in Machine Learning. There are plenty of methods one can utilize for better understanding the behaviour of a model. Here we approach the explanation of regression prediction problems through the usage of visual methods, since these are more adequate for conveying information to end users with reduced technical background.

While most existing work analyses the output of the model, we claim that assessing the performance of the model is also of high relevance. The contributions of our work are then more focused on this latter aspect. We develop a novel approach to inspect the estimated risks of using a black-box regression model for a concrete test case. We describe, evaluate and propose tools that visually convey the relationship between the expected error and the values of a predictor variable.

Moreover, we address a real-world problem, in which we our tools and other state-of-the-art methods to understand factors that drive fishing effort around Large Scale Marine Protected Areas. Additionally, we compare some predictive algorithms to select the most suitable for the problem and then provide an overview analysis of the performance of the chosen one.

Keywords: Explainability, Black box model, Regression, Performance, Accountability, Large Scale Marine Protected Areas

Contents

List of Tables

List of Figures

Nomenclature

Roman Symbols

f	Function.
$\hat{f}$	Prediction model function.
N	Number of instances of training set.
X	Predictor variable.
Y	True target variable.
$\hat{Y}$	Predicted target variable.

Subscripts

i	Instance index.

Superscripts

C	Index of complement of features of interest.
p	Predictor index.
S	Features of interest index.

Acronyms

AI Artificial Intelligence. 2, 4

AIS Automatic Identification System. 44

ALE Accumulated Local Effects. 14, 15, 54

CV Cross Validation. 22, 47

d-ICE Derivative Individual Conditional Expectation. 13

DARPA Defense Advanced Research Project Agency. 2

EDP Error Dependence Plot. 22, 23, 30, 50, 67, 68

EEZ Exclusive Economic Zone. 45, 46, 57

FD Finite Difference. 14

GBM Gradient Boosting Machine. 24, 26, 27, 36

ICE Individual Conditional Expectation. 12, 13, 17

ICI Individual Conditional Importance. 18

KS Kolmogorov-Smirnov. 30, 31

LIME Local Interpretable Model-Agnostic Explanations. 17

LIVE Local Interpretable Visual Explanations. 17

LSMPA Large-Scale Marine Protected Areas. 2, 3, 8, 43, 44, 52, 68

MARS Multivariate Adaptive Regression Spline. 6, 46, 47, 68

MCR Model Class Reliance. 11

MEDP Multiple model Error Dependence Plot. 37, 47, 50, 67

ML Machine Learning. 1, 5, 67

MNM Marine National Monument. 80

MP Marine Park. 80

MPA Marine Protected Areas. 43, 44, 46, 80

MR Model Reliance. 11

MR Marine Reserve. 46, 80

NN Neural Network. 29, 37

OOB Out-Of-Bag. 6

PA Protected Area. 80

PDP Partial Dependence Plot. 11–13

PEP Parallel Error Plot. 34, 50, 67

PFI Permutation Feature Importance. 10, 11, 18

PI Partial Importance. 18

QII Quantitative Input Influence. 18

RF Random Forest. 6, 35, 46, 47, 68

SHAP SHapley Additive exPlanations. 17, 18

SST Sea Surface Temperature. 44, 47, 53, 61

SVM Support Vector Machine. 27, 46, 47, 68

VINE Visual INteraction Effects. 17

XAI eXplainable Artificial Intelligence. 2, 6, 10

Chapter 1

Introduction

1.1 Motivation

A demand for transparency

Machines have great advantages when comparing to humans in terms of reproducibility, scaling, speed and, with the evolution of algorithms, even accuracy. Nevertheless, sophisticated machine learning algorithms developed recently reached a complexity level that inherently hinders their functioning. These models are often named as black-boxes or opaque models, meaning that one can neither understand their internals nor the reasons behind certain predictions.

As these models begin driving important and costly decisions, for instance related to human health or finances, end users have been pressuring for explainability and transparency. In fact, as much accurate as the model might be, decision makers will always require an explanation in order to fully trust it.

Gilpin et al. [1] defined explainability as the ability 'to summarize the reasons for [the model] behavior, gain the trust of users, or produce insights about the causes of their decisions'. The lack of this characteristic was proven to be problematic in several circumstances, having lead to cases of predictions being based on the wrong factors [2, 3] or even gender bias [4] and race bias [5, 6].

Burrell [7] identified three types of barriers to transparency:

1. intentional secrecy on the part of the institutions, to avoid public scrutiny;

2. lack of technical literacy from the part of the end user, implying that even accessing the code would not be sufficient to understand the model;

3. the complexity of machine learning models, with high-dimensional characteristics.

The focus of this work is in respect to the last barrier. So far, the tendency in Machine Learning (ML) has been focused on the improvement of performance and accuracy, with consequent increase in the complexity of the algorithms formulated. However, having only that in consideration has led to the production of more and more opaque models, as for example Random Forests and Neural Networks. In fact, advanced ML algorithms tend to be non-interpretable due to the fact that their complexity diverges from the human-level of reasoning and interpretation capabilities. Thus, users have to decide whether to opt for an interpretable model in exchange for lower accuracy or for a highly accurate but opaque model.

This compromise is one of the largest obstacles to the wide-scale adoption of machine learning. In this context, implementing explainable models and understanding black-box models has become one of the hot topics in Artificial Intelligence (AI) research, renewing attention on the eXplainable Artificial Intelligence (XAI) branch. More specifically, according to DARPA [8], the aim of XAI is to 'produce more explainable models, while maintaining a high level of learning performance (prediction accuracy); and enable human users to understand, appropriately, trust, and effectively manage the emerging generation of artificially intelligent partners'.

It is widely acknowledged that the need for transparent and fair algorithms is urgent, this being one current great challenge in Machine Learning and Data Science. This demand has been pushed even further due to the EU General Data Protection Regulation (GDPR), that took effect in May 2018. This law leveraged the debate of the social right to explanation, which states that an individual has the right of explanation in the case of an automated decision.

The aerospace sector is no exception to this urgency, since predictive modelling techniques are currently being used and investigated for purposes of failure prediction, development of intelligent avionics systems, supply chain management, CFD analysis, optimized flight performance, air travel demand prediction and research of new materials. All these different domains require complex models, that inevitably act as black-boxes. Hence, it is necessary the development of explainability mechanisms that dissect these same models.

There are plenty of methods one can utilize for better understanding the behaviour of a model. In this work, we focus on the explanation of regression prediction problems through the usage of visual methods, since these are more adequate for conveying information to end users with reduced technical background.

Most existing work on XAI analyses the output (predictions) of the algorithm. While this is very important for understanding the model, we claim that assessing the performance of the model is also of high relevance, particularly when the predictions drive costly decisions. The contributions of our work are then more focused on this second aspect of explainability. Nevertheless, we will also address a real world application where we will use both methods (analysing the output as well as explaining the performance) to fully explain and understand black-box models.

Large Scale Marine Protected Areas and Global Fishing Fleets

As a motivation, we aim to study the fishing effort on Large-Scale Marine Protected Areas (LSMPA). LSMPAs encompass marine areas over $100000km^2$ designated to enhance marine ecosystems protection and resist threats originated by overfishing, climate change and coastal development. However, their quite recent establishment entails a scarcity of information on the interactions between LSMPAs and surrounding fisheries. Thus, experts are requiring insight into the factors that influence this fishing effort in order to better drive sustainability policy decisions on these important LSMPAs.

We propose using a data set provided by experts to model the fishing effort of ships around different areas of the globe that are neighbours to Large-Scale Marine Protected Areas (LSMPA) and thereupon employ explainability tools in the resultant models to understand which factors drive fishing effort.

1.2 Objectives and Contributions

This thesis proposes an comprehensive study of the state-of-the-art methods that help explaining black-box regression models, with a particular focus on visual tools. We aim at differentiating between tools that interpret how each variable influences the model prediction and tools that explain and predict the performance of the model.

We identify some gaps in the latter methods, which do not present a solution for predicting the performance of the model for specific domain conditions. Hence, we develop three visual tools and respective variants that serve this purpose. Under this topic, two publications were presented and published in the respective conference proceedings: *Visual Interpretation of Regression Error* [9] at the 19th EPIA Conference on Artificial Intelligence and *Explaining the Performance of Black-Box Models* [10] at the 6th IEEE International Conference on Data Science and Advanced Analytics. The former was recently selected by the conference organizers for submission of an extended version to the journal Expert Systems while the latter was invited to be extended to the Journal of Data Science and Analytics, both currently under preparation. Furthermore, the developed software in this scope is openly available and can be found in `http://github.com/inesareosa/MScThesis/Performance`.

Finally, we implement the most adequate methods to analyse the fishing effort of vessels around different areas of the globe near to Large-Scale Marine Protected Areas (LSMPA). The main goal is to provide useful information for an end user on the factors that influence the fishing effort. Using the previously developed tools, we begin by comparing predictive algorithms to select the most suitable for the problem, following with an overview analysis of the performance for the chosen algorithm - a Random Forest. Lastly, we employ some selected state-of-the-art interpretability tools to fully understand the impact of a set of environmental, physical and economic factors on the fishing effort.

1.3 Thesis Outline

In the present chapter the work is introduced concisely, defining the motivations for the thesis as well as the contributions.

In Chapter 2 the problem is defined and the objectives are formulated. Some core concepts necessary for the understanding of the subject are clarified and the data sets and the predictive models used throughout the work for benchmarking are introduced. Hereafter, the two paradigms in explainability are well distinguished and dissected according to the existent state-of-the-art methods.

Chapter 3 addresses the evaluation of the performance of black-box regression models. Previously existing methods are found to focus only on the overall global performance or on the performance for different operating conditions. Hence, an original perspective for analysing the performance of a single or multiple models is here introduced with the proposal of three novel tools, studied and evaluated using standard black-box regression tools.

In Chapter 4 a case study that relates Large Marine Protected Areas and fishing effort by global fleets is defined. Considering the methods proposed in Chapter 3 and the tools already existent introduced in Chapter 2, the models that study this relation are analysed and assessed. A comparative study on which machine learning tool is more suitable for the problem is also performed.

Lastly, in Chapter 5 we provide an overview of the main conclusions and contributions of this work within the field of Explainable AI, outlining possible areas for future research and development of this work.

Chapter 2

Background

2.1 Core Concepts

Machine Learning (ML) is a multidisciplinary field that aims to create machines that automatically learn how to solve new talks from experience, without being programmed to do so [11]. One of the many tasks addressed by ML focuses on predictive analysis (supervised learning), which seek to learn associations and relations from complex data, in order to be able to predict an output based on previous experience.

The complex data used to acquire information can assume various types, such as categorical, nominal, text, time series, audio, video or images. Tabular data refers to a data set in which the observations are described by a fixed set of properties, usually conveyed by a table with numeric or/and categorical data, containing information on the features and the target value. The predictors (attributes, input variables, input features) are the inputs for the prediction algorithms, while the target (dependent variable, target feature) is the value that the model aims to predict. A prediction is the value that the model assumes as the most adequate for a set of input features. The data is denominated as training data if used as input of the model and as test data if used to test the validity of the model predictions. An training instance consists of the feature values and the respective target outcome for a single observation, corresponding to a row in the data set.

Moreover, the type of the target variable defines the type of the task to be addressed: the main ones being either classification or regression. Regression encompasses tasks that attempt to predict a quantitative response, while classification regards tasks that intend to predict a categorical response.

Correlated features refer to variables that may have some form of dependency with other variables, either positive or negative. The correlation between two features can be calculated using the Pearson or the Spearman correlation coefficient [12, 13]. The Pearson coefficient should be chosen if both features have a normal distribution and with the intent of discovering linear relationships, while the Spearman should be used to investigate monotonic relationships. When two features are correlated, certain combinations of those features might not be feasible to occur in the real world. Assuming the training data only covers a limited range of domain, we make an extrapolation whenever we predict in unseen domains (outside of the training data distribution).

A variable interaction describes the effect of the response function that is originated by the relationship between more than one feature, in which the effect of a variable depends on other variables. This implies that in cases in which such interactions are present, the response function cannot be considered as a sum of independent variable effects.

Predictive performance aims to evaluate the agreement between the true label of an instance provided in the data set, also denominated as ground truth, and the actual prediction outputted by the model.

Throughout this work we will apply various machine learning algorithms, from Random Forests [14] to Support Vector Machines [15], Gradient Boosting Machines [16], Neural Networks [17] and Multivariate Adaptive Regression Splines [18]. All the models are used without much depth, just as a tool to analyse after being trained. However, throughout some parts of the work it is taken a special focus on Random Forest (RF) and on some inherent concepts. In a nutshell, RFs generate an ensemble of a large number of decision trees and average the results obtained in each tree, using *bagging* techniques as well as a randomly selected subset of features in each tree split point to avoid correlation between them. Bagging, or Bootstrap Aggregation, is the process of each individual tree being trained using only a random subset of instances of the original data set.

Out-of-bag (OOB) refers to the data that was not used when constructing each tree. Node purity is a measure of the homogeneity of the splitting node, that checks to which extent does the model split well the data. For regression, the metric most commonly used is the residual[1] sum of squares within the node [20].

2.2 Problem Formulation

A black-box model is defined as a system that does not disclose its internal mechanisms. This implies that an end user cannot understand the model even if looking at the parameters, and consequently cannot apprehend whether the algorithm is acting as supposed. A white box, or transparent box, is the exact opposite, and refers to an intrinsically interpretable model.

Most explainability tools proposed recently try to explain the outcome of a model for certain predictors conditions, answering to the problem of interpretability. These methods intend to grasp cause and effect phenomenons, trying to understand the influence of the inputs in the predicted outcome. However, in order to trust a prediction it is crucial to provide an assessment of the risk of the model, and on this necessity lies the problem of accountability.

To address explainability problems, a new field of research is emerging: the explainable AI (XAI), which aims to develop methods to interpret and evaluate black-box models, in order to further comprehend the functioning of the system and consequently understand the decisions behind certain predictions. The explanations can be provided through a multitude of means, such as visual aids or textual reasoning. Visualization is more convenient for compiling complex ideas and summarizing relations be-

[1] measure of disagreement between data and assumed model [19]

tween inputs, while textual explanations have the advantage of transmitting information in human-like manner and of being straightforward. As we are interested in capturing complex relations, the focus of this work will be on studying and developing visual methods, both regarding interpretability and accountability problems.

Supervised machine learning techniques can serve two distinct types of tasks: regression or classification, depending on the type of the target variable. This work prioritizes regression models, since explainability within classification problems has been extensively studied [21], specially in terms of performance analysis [22–24].

The explainability tools can be divided into model specific or model agnostic ones. The former relates to methods that are specific to a certain algorithm, only working for the interpretation of a certain class of models. On the other hand, model agnostic methods comprise tools that can be applied independently of the type of algorithm, usually by analysing the relation between the input and the respective prediction, without the need for observing into the model internals.

Interpretability can be achieved through intrinsic or *post-hoc* methods. Intrinsic methods, also named as transparent box design or *ante-hoc* methods, refer to cases in which the model was already created to be interpretable on its own [25], with a likely consequent decrease of complexity and performance. *Post-hoc* methods refer to the application of tools that analyse a pre-trained model, complex or not [26].

Our proposal intends to study *post hoc* visual tools that help explaining black-box regression models, with a priority for model-agnostic techniques. The methods in study approach problems of accountability and interpretability, with a greater focus on the former, in which we propose to develop novel tools.

2.2.1 Notation

Concerning the notation used throughout the work, assume a p-dimensional predictor space $\mathcal{X}^p$ and a target space $\mathcal{Y}$. The function f identifies the unknown relationship between $\mathcal{X}^p$ and $\mathcal{Y}$. The function $\hat{f}$ represents the obtained prediction model that approximates f, to be analyzed with the *post-hoc* interpretability and accountability methods. The predictors variables are represented by $\mathbf{X} = (X^1, ..., X^p) \in \mathcal{X}^p$, whereas $Y \in \mathcal{Y}$ represents the true target value and $\hat{Y} \in \mathcal{Y}$ represents the predicted variable ($\hat{Y} = \hat{f}(\mathbf{X})$). The regression model is trained with data $\{\mathbf{x_i}, y_i\}_{i=1}^N$, where $\mathbf{x_i} = (x_i^1, ..., x_i^p) \in \mathcal{X}^p$ is a vector of predictors for a single instance, $y_i \in \mathcal{Y}$ is the true target value for that same observation and N is the number of observations in the training set.

Let $S \subset \{1, ..., p\}$ and let C be the complement set of S, both representing subsets of predictors indexes. Thus, $\mathbf{X}^S$ represents a target subset of the predictor variables $\mathbf{X}$, known as features of interest, and $\mathbf{X}^C$ represent the complement subset ($\mathbf{X} = \mathbf{X}^S \cup \mathbf{X}^C$).

2.2.2 Tools and Material

All the experiments were conducted using R programming language [27], thus enabling the full reproducibility of the analysis and subsequent results. The used R packages are accredit in each corresponding section. All plots and figures in this dissertation were generated using the R package *ggplot* [28].

Data Set	*Inst*	*Pred*	*Data Set*	*Inst*	*Pred*
a1	198	11	a2	198	11
a3	198	11	a4	198	11
a6	198	11	a7	198	11
Abalone	4177	8	acceleration	1732	14
availPwr	1802	15	bank8FM	4499	8
cpuSm	8192	12	fuelCons	1764	37
boston	506	13	maxTorque	1802	32
servo	167	4	airfoild	1503	5
concreteStrength	1030	8	machineCpu	209	6

Table 2.1: Data sets used for benchmarking ($Inst$: number of instances; $Pred$: number of predictors).

Learner	Parameters	R package
NN	$size = 10, decay = 0.1, maxit = 1000$	**nnet** [17]
SVM	$cost = 10, gamma = 0.01$	**e1071** [29]
RF	$ntree = 1000$	**randomForest** [30]
GBM	$distribution = ”gaussian”, n.trees = 5000,$ $interaction.depth = 3$	**gbm** [16]

Table 2.2: Regression algorithms, parameters, and respective R packages used for the benchmarking.

Benchmark Data Sets and Models

For enabling the test of the tools developed and presented in Chapter 3 while allowing replicability, we used a set of 18 real world data sets from different domains, with variable size and number of predictors, as defined in Table 2.1. These are publicly available in `https://github.com/inesareosa/MScThesis/Datasets`.

Each data set was modelled as a regression task using the predictive learning algorithms described in Table 2.2. The diversity of models selected (Random Forest, Neural Network, Support Vector Machine and Gradient Boosting Machine) avoids the existence of model-dependent bias within the experiment conclusions.

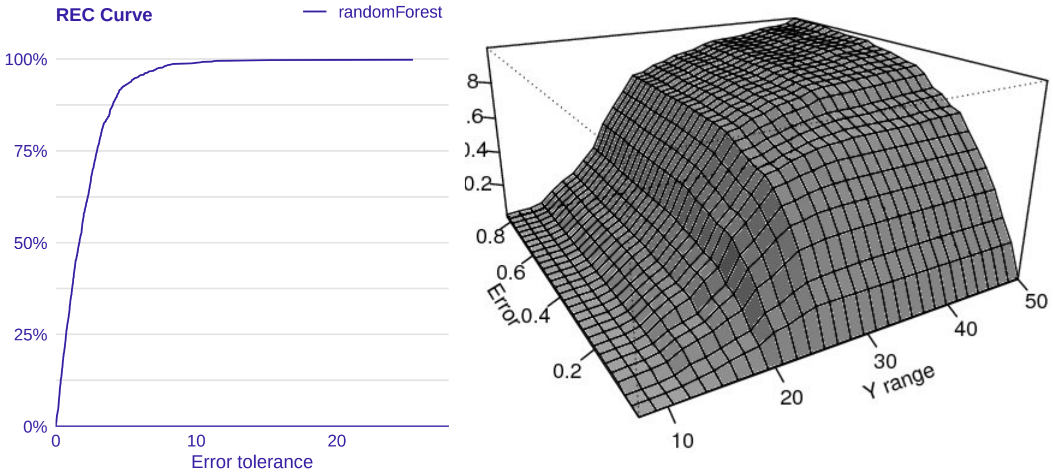
Fishing Effort Data Set

The data set for the proposed case study was provided by Dr Kristina Boerder. This contains information of fishing effort inside and within 500km from the border of 13 Large-Scale Marine Protected Areas (LSMPA). It is also provided input on the characteristics for each element in the spacial grid, as well as characteristics related to each particular LSMPA. Additional detailed information is provided in Section 4.2.

2.3 Assessing the Black-Box Performance

The evaluation of a regression model hinges on the differences between the true and predicted values - the prediction errors, and can be executed using scalar or graphical metrics. The former method, most commonly used, quantifies an estimate of the expected error, using approaches such as the Mean Squared Error, the Root Mean Squared Error, the Mean Absolute Percentage Error and the Median Error [31, 32]. However, this method provides a single metric for the entire model, concealing information if certain predicted values tend to be more error prone. Other methods, such as the Akaike Information Criterion [33] and the Bayesian Information Criterion (BIC) [34] compare the quality of several models by estimating the relative loss of information, to balance overfitting and underfitting, using the number of parameters (and the number of observations for BIC) as a measure of complexity. However, once again these methods only take into consideration the overall model.

Graphic metric approaches provide a different perspective on the analysis of the model, informing about the changes in the performance for different operating conditions, as are examples lift charts [35], RROC space [36], REC curves [37] and REC surfaces [38]. REC curves, with an example depicted in Figure 2.1, plot the error tolerance versus the percentage of points predicted under that same tolerance, representing an estimation of the error cumulative distribution function of a model. REC surfaces, exemplified in Figure 2.1, in turn, add the target values to this graphic.



(a) REC plot from data set *cementStrength* (c.f.Table 2.1) trained with a Random Forest

(b) REC Surface (taken from [38])

Figure 2.1: Graphical performance assessment methods.

Existing tools only address the error or the error tolerance in respect to the target value. These methods assess the model as an whole and do not consider that different conditions might lead to distinct performance behaviours. From our research, we could not find in the literature any method that establishes a relationship between the (expected) error and the predictor variable values, a perspective that we consider important since it would provide an explanation for certain error patterns.

2.4 Interpreting the Black-Box

In the previous section we have reviewed efforts on trying to understand the predictive performance of the models. This section focuses on methods that try to explain the value predicted by a black-box model instead of its prediction error. As recent regression machine learning algorithms tend to be intrinsically complex, these do not inform the end user about the reasoning behind a certain prediction. Hence, the aim of interpretability is to enhance the knowledge on how the calculation of the target value is influenced by each predictor.

The main innovations and endeavours in XAI can actually be represented by the considerable number of interpretability tools that have been being suggested recently. The present section dissects the most well established *post-hoc* methods for interpreting a black-box regression model, as well as some cutting edge tools developed recently, with a prioritization of model agnostic techniques. This means we will mainly describe tools that scrutinize any type of regression algorithm after it has been trained. Thus, all the methods presented require the learned model as input, with some necessitating other information, such as the training data, including or not the true predictions.

The choice of the adequate explanation methods strictly relies on the problem context, being also influenced by time limitations as well as by the personal preferences of the users, their perception of the real world scenario reflected by the model and their technical literacy.

Two paradigms prevail in the interpretability techniques, distinguishing the methods between global and local explanations. Global methods intend to describe the functioning of a model in broad terms. On the other hand, local methods concern the explanation of the prediction for a specific instance, chosen by an end user. Recently, the concept of regional explanations has been introduced by Britton [39], defining methods that aim to interpret a set of instances, especially when these have a behaviour that differs from the global explanation.

Bear in mind that inherently interpretable models are an example of global explanations. However, we will not approach them since these are not *post-hoc* methods. These models comprehend decision trees, linear regressions or rule lists, and can be interpreted by simply observing the resultant model.

2.4.1 Global Methods

2.4.1.1 Feature Importance

A straightforward assessment of the model can be performed by inspecting the relative influence of a feature X^p on the prediction function $\hat{f}$, through a certain feature score. Several criteria have been proposed for regression algorithms such as Gini Importance [40], Regressional ReliefF [41], Permutation Feature Importance (PFI) [14], Model Class Reliance [42] and Shapley Feature Importance [43].

PFI and Gini Importance are model-specific methods, both employed in Random Forests. The latter, also known as Mean Decrease in Impurity or Mean Increase in Purity indicates the number of times that the variable is used for splitting a node averaged over all trees, measuring the homogeneity of the feature.

The PFI estimates the importance of a given feature by computing the increase in the prediction error when the values of that feature are permuted. This is based on the perception that if a feature is not important for obtaining a prediction, then permuting its values will not reduce significantly the performance and accuracy of the model. The decrease in the predictive accuracy due to the permutation of a single feature is calculated using the out-of-bag samples of each tree, frequently utilizing the mean squared error as the performance metric. The values obtained are then averaged over all trees.

PFI is generally considered to be an efficient technique, but the existence of correlated features might decrease the importance of each one by dividing the importance between both. Furthermore, the correlation of features can lead to the creation of unrealistic data instances upon the permutation, that will consequently drive the PFI into being calculated with values that are not feasible in reality.

Fisher et al. [42] developed the PFI method into a model-agnostic version: the Model Reliance (MR), suggesting two methods of permutation: either splitting the data set into two folds and exchanging the values of the variable of the two groups, or permuting over all possible ($n!$) combinations, with n being the number of rows in the data set. The pseudocode in Algorithm 1 describes the process of calculating the MR in more detail. The authors further develop this metric, suggesting the Model Class Reliance (MCR), which indicates the highest and lowest degree to which a certain class of models depends on the features of interest to predict well, using the permutation-based importance estimates.

Algorithm 1: Pseudocode for obtaining the Feature Importance using the MR method.

```
input  : predictors data $\mathbf{X}$
input  : target value data $\mathbf{y}$
input  : trained model $\hat{f}$
input  : performance metric $Err$ (e.g. Mean Squared Error)
input  : permutation method $Permute$
output: model reliance $I$

$\hat{e}_{original} \leftarrow$ Err($\mathbf{y}$, $\hat{f}(\mathbf{X})$)
$p \leftarrow$ NumberofPredictors($\mathbf{X}$)
$I \leftarrow \{\}$

foreach $k$ in $\{1,...,p\}$ do
    $\mathbf{X}_{perm} \leftarrow$ Permute($\mathbf{X}, k$) // permute values of predictor $X^k$, all others remain the
        same
    $\hat{e}^k_{mr} \leftarrow$ Err($\mathbf{y}$, $\hat{f}(\mathbf{X}_{perm})$)
    $I^k \leftarrow \frac{\hat{e}_{mr}}{\hat{e}_{original}}$
return $I$
```

Both the PFI and the MR take into account not only the importance of the feature itself but also the importance related to interaction effect with other features. However, the permutation methods depend on the randomness of the shuffling, meaning that the results might vary and that the operation should be repeated to assure reliable results. A more accurate method can be performed with the drop-column importance [44], which calculates the decrease in the performance when the feature is removed. Nevertheless, this method presents an higher computationally time when comparing to PFI and MR, as it requires the retraining of the whole model without the predictor.

Another critic to the presented feature importance techniques lies on the fact that these are linked to estimated error and not actually to the output itself.

2.4.1.2 Feature Effect Plots

Proposed by Friedman [45], the Partial Dependence Plot (PDP) maps the average predicted model outcome across the range of values of a feature of interest, as depicted in Figure 2.2. In fact, PDPs were the inspiration for our proposal of EDPs, described in Section 3.2.1, which instead of relating the model performance with the predictor variables values, uses the model performance.

PDPs estimate the output of a prediction for each value of a certain feature of interest X^S, computing the average predicted value $\hat{f}$ when X^S is fixed and $\mathbf{X^C}$, the complement features, changes over its marginal distribution:

$$\hat{f}_{PDP}(x^S) = \mathbb{E}_{\mathbf{x_C}}[\hat{f}(x_S, \mathbf{x_C})] = \int \hat{f}(x_S, \mathbf{x_C}) P(\mathbf{x_C}) d\mathbf{x_C}, \tag{2.1}$$

with $P(\mathbf{x_C})$ representing the probability of occurrence of $\mathbf{x_C}$. Equation 2.1 is estimated using the training data with n instances: for each grid value of $X^S = v_k$, n cases are forged ($< v_k, \mathbf{X}^C >$), with $\{\mathbf{X}_1^C, ..., \mathbf{X}_n^C\}$ designating the values of the predictors other than the feature of interest observed in the training data. This is, each instance in the data set is merged with the feature of interest of value v_k. The black-box model is then queried with these cases, obtaining n predictions that should be averaged to provide $\hat{f}(v_k)$. This process is then repeated across the desired grid values to obtain the full plot, as described in Equation 2.2.

$$\hat{f}_{PDP}(x^S) = \frac{1}{n} \sum_{i=1}^{n} \hat{f}(x^S, \mathbf{X}_i^C), \tag{2.2}$$

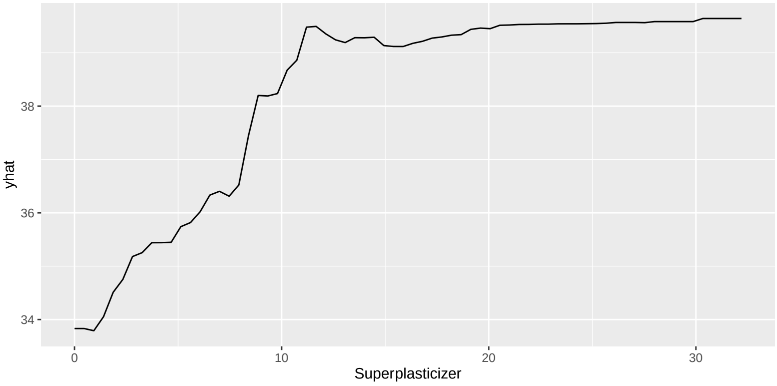


Figure 2.2: PDP of feature *Superplasticizer* from data set *cementStrength* (c.f.Table 2.1) trained with a Random Forest (obtained using R package *pdp* [46]. The PDP shows that the target value increases with the the increase of amount of *Superplasticizer*, stabilizing for high (>15) values.

PDPs even leveraged other methods to analyse feature importance, as proposed by Greenwell et al. [47], which uses the flatness of the PDP as a metric for evaluating the importance of the feature.

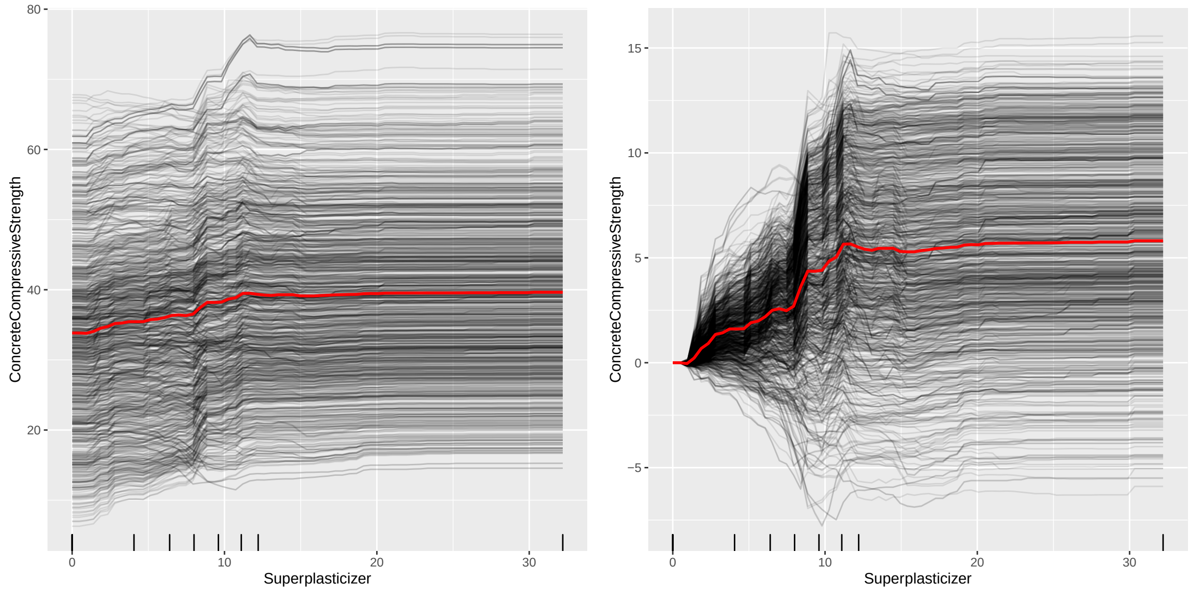
The partial dependence function is of easy interpretation for an end user, but the averaging function might hide heterogeneous effects caused by feature interactions.

In order to disaggregate the PDP, revealing the full complexity of the model, Goldstein et al. [48] suggested the Individual Conditional Expectation (ICE) plots, which display the n estimated curves (one for each set of values in $\mathbf{X}^C$ observed in the training data) instead of the average partial effect. Each curve defines the conditional relationship between x^S and $\hat{f}$ at the values of observation $\mathbf{X}_i^C$.

Nevertheless, ICE do not scale well with the number of data instances, as more data cases imply an increasing number of lines in the plot and consequent overploting. Centered ICE and Derivative ICE plots (d-ICE) explore the presence of interaction effects, while uncluttering the data display. The former removes level effects, as shown in Figure 2.3, displaying only the cumulative effects:

$$\hat{f}_{PDPcentered_i} = \hat{f}_i - \hat{f}(\mathbf{x}^*, \mathbf{X}_i^C) \tag{2.3}$$

in which $\mathbf{x}^*$ represents the location in range of X^S where the prediction lines are ought to be joined, usually the minimum value of X^S, ensuring all curves begin at $\hat{f}_{centered} = 0$.



(a) Regular ICE plot

(b) Centered ICE plot

Figure 2.3: ICE plot of feature *Superplasticizer* from data set *cementStrength* (c.f.Table 2.1) trained with a Random Forest (obtained using R package *ICEbox* [49]. In red is plotted the corresponding PDP. The marks in the X-axis mark the 10% quantiles of distribution. In the left plot we can observe two main patterns, confirmed when we remove the levels in the centered version, indicating the existence of an interaction with other features.

Derivative ICE plots (d-ICE), on the other hand, explore interactions through the derivation of the PDP in terms of the feature of interest. Considering a case in which X^S does not interact with the other features, then we would have

$$\hat{f}(\mathbf{x}) = \hat{f}(x^S, \mathbf{x}^{\mathbf{C}}) = g(x^S) + h(\mathbf{x}^{\mathbf{C}}) \tag{2.4}$$

$$\frac{\partial \hat{f}(\mathbf{x})}{\partial x^S} = g'(x^S). \tag{2.5}$$

Equation 2.5 then implies that, in the absence of interactions, the n lines in the d-ICE plot should be similar. Observing the d-ICE plot will then allow the end user to analyse whether the model presents more than one behavior of the derivative curve, as it is the case in Figure 2.4, which would indicate the presence of interactions.

For cases in which the features in X^S are correlated with $\mathbf{X}^C$, the calculation of both PDP and ICE will wield unlikely data points, since it fabricates instances that might not be possible in the real world, which will consequently lead to the creation of erroneous plots [50].

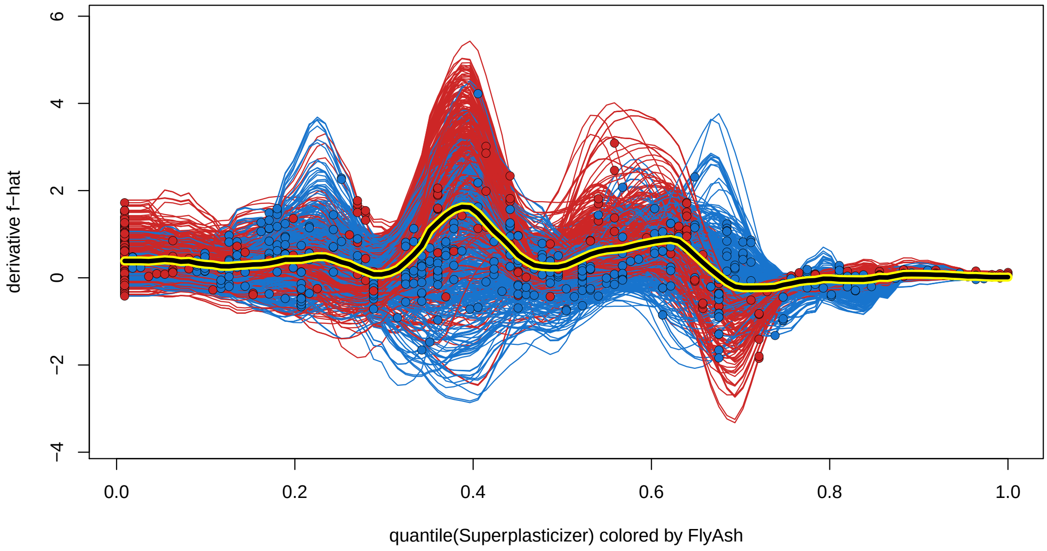


Figure 2.4: d-ICE plot of quantiles of feature *Superplasticizer* from data set *cementStrength* (c.f.Table 2.1) trained with a Random Forest, coloured by values of *FlyAsh* (obtained using R packages *ICEbox* [49]. In yellow is plotted the average d-ICE. The different behaviours of the lines in the d-ICE plot show the presence of interactions. In fact, when plotting different colors for *FlyAsh* (blue if higher than 50 and red if lower), we can notice that this feature is probably interacts, since similar values have similar patterns.

Accumulated Local Effects (ALE), exemplified in Figure 2.5, introduce a new approach to visualize main effects, avoiding extrapolations due to correlated features while being less computationally expensive than the previously mentioned plots [50]. This method calculates the changes in the model predictions for data instances in small windows of the feature of interest, not requiring the synthesis of new data.

Having $\hat{f}^S(x^S, \mathbf{x^C}) = \frac{\partial \hat{f}(x^S, \mathbf{x^C})}{\partial \mathbf{x^S}}$, the ALE main effect for a predictor is given by:

$$\begin{aligned}\hat{f}_{ALE}(x^S) &= \int_{z_o^S}^{x^S} \mathbb{E}[\hat{f}^S(X^S, \mathbf{X}^C)|X^S = z^S]dz^S - constant \\ &= \int_{z_o^S}^{x^S} \int_{\mathbf{x}^C} \hat{f}^S(z^S, \mathbf{x}^C)P(\mathbf{x}^C|z^S)d\mathbf{x}^C dz^S - constant,\end{aligned} \tag{2.6}$$

with z_o^S being a value near X^S lower bound and the constant being calculated to center the plot vertically.

To estimate the first-order ALE, as most models do not present information on the gradient, the domain of X^S is partitioned into a grid with K intervals. For each observation within an interval, we compute the prediction finite difference by setting the value of X^S as the upper and lower boundary of that same interval:

$$FD = \hat{f}(X^S = upperlimit, \mathbf{X}_i^C) - \hat{f}(X^S = lowerlimit, \mathbf{X}_i^C), \tag{2.7}$$

obtaining what is called as the *Effect*, which corresponds to the derivative $\hat{f}^S(X^S, \mathbf{X}^C)$. The FDs inside each interval are then averaged - providing the *Local Effects*, which corresponds to the expectation

$\mathbb{E}[\hat{f}^S(X^S, \mathbf{X}^C)|X^S = z^S]$. Lastly, the expectation of all intervals is integrated by summing the average values obtained in each window - and here we have the *Accumulated Local Effects*:

$$\begin{aligned} \hat{f}_{ALE}(x^S) &= \hat{f}_{uncentered}(x^S) - constant \\ &= \sum_{k=1}^{k(x^S)} \frac{1}{n^S(k)} \sum_{i:x_i^S \in N^S(k)} [\hat{f}(z_k^S, \mathbf{x}_i^C) - \hat{f}(z_{k-1}^S, \mathbf{x}_i^C)] - constant, \end{aligned} \tag{2.8}$$

in which $N^S(k)$ indicates the interval between the lower and upper limits $]z_{k-1}^S, z_k^S]$, $n^S(k)$ indicates the number of instances in the interval k and with the constant being estimated as:

$$constant = \frac{1}{n} \sum_{k=1}^{K} n^S(k) \hat{f}_{uncentered}(z_k^S). \tag{2.9}$$

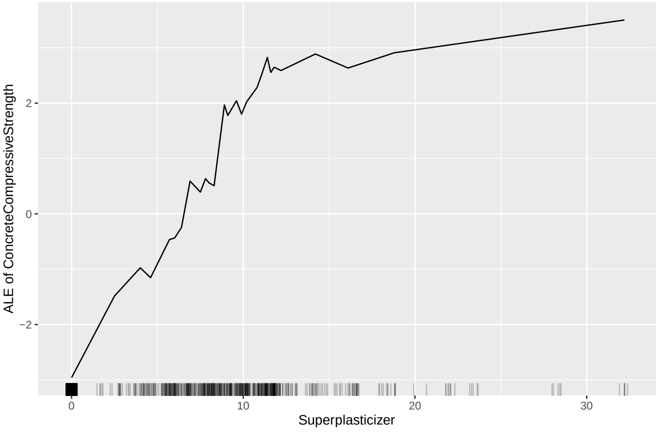
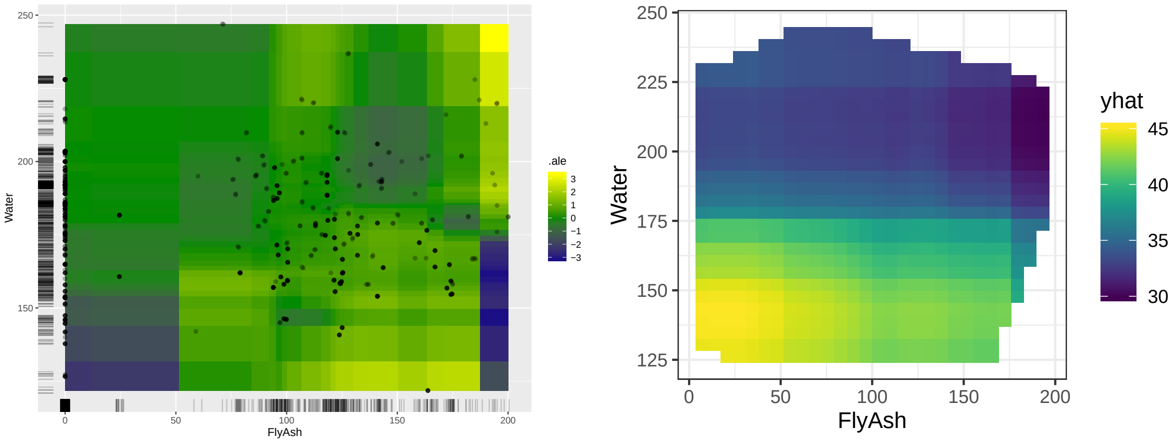


Figure 2.5: ALE plot of quantiles of feature *Superplasticizer* from data set *cementStrength* (c.f.Table 2.1) trained with a Random Forest (obtained using R package *iml* [51]). The traces in the X-axis represent the values of the feature found in the training set.

2.4.1.3 Feature Interaction

Besides exploring the effects that each feature has on the predicted outcome, it is crucial to probe more complex effects, namely prominent interactions between features, also denominated as non-additive interactions.

Both ALE plots and PDPs present neat variants that visually relate the expected prediction to the domain values of 2 features: the second-order ALE plots and the two dimensional PDPs [45, 50]. An example of both is plotted in Figure 2.6. While 2D PDPs account for the total effect of the two variables ($PDP(x^j, x^k) = f(x^j) + f(x^k) + f(x^j, x^k)$), 2D ALE plots only provide the visualization of the additional interaction between two features - the second-order effect: ($ALE(x^j, x^k) = f(x^j, x^k)$). The former can be calculated by substituting x^S by $\mathbf{x^S}$, a vector of two predictor variables, in Equations 2.1 and 2.2. The latter uses similar calculations as with one feature, but utilizes K^2 rectangular cells instead of K intervals in a grid, accumulating the local effects in two dimensions.



(a) Second-order ALE plot

(b) 2D PDP

Figure 2.6: Interaction plot between features *FlyAsh* and *Water* of data set *cementStrength* (c.f. Table 2.1) trained with a Random Forest (obtained using R packages *iml* [51] and *pdp* [46]). Note that the ALE plot (on the left) plots only the interaction effect, while the 2D PDP plots the total effect of both figures.

In order to actually measure feature interactions and obtain a score, some methods based in some principles of the 2D PDP were developed by Greenwell et al. [47] and Friedman and Popescu [52]. The latter, named H-Statistic, is quite computationally expensive but can detect all types of interactions. This metric generates a score from 0 to 1 for each feature pair, with 0 indicating the absence of any interaction between the features and with 1 representing that the effect on the prediction is solely generated from their interaction.

H-Statistic is based on the assumption presented in Equation 2.4, which states that if two variables do not interact, then the partial dependence can be decomposed into the sum of each variable partial dependence function. The metric then measures the fraction of the variance of the two dimensional partial dependence function that is not captured by the sum of the each variable partial dependence function:

$$H^{jk} = \frac{\sum_{i=1}^{n}[\hat{f}_{PDP}(x_i^j, x_i^k) - \hat{f}_{PDP}(x_i^j) - \hat{f}_{PDP}(x_i^k)]^2}{\sum_{i=1}^{n}\hat{f}_{PDP}^2(x_i^j, x_i^k)} \tag{2.10}$$

While this method is efficient in clearly identifying the terms with more interaction, it does not pinpoint which regions in the domain present the strongest interactions.

VIN [53] graphically identifies interactions between all features, including the ones involving 3 or more terms. However, it does not indicate for which cases or values the interactions occur, neither the strength of the interaction.

2.4.1.4 Global Surrogate

A global surrogate is an interpretable model that emulates the original black-box model, usually trained with data resultant from querying the machine. The interpretable models can range from decision trees [54–56] to general additive models or decision rules [57].

2.4.2 Regional Explanations

Britton [39] introduces the novel concept of regionality, addressing groups of instances that act as exceptions to the global model behaviour. The specific tool proposed, VINE, tries to highlight distinct groups by clustering ICE curves according to their slopes. A decision tree is then trained to generate an explanation that distinguishes each group from all other instances. These groups are then merged if having similar explanations.

However innovative this approach might be, we believe it should only be used in a pre-analysis phase to automatically explore some interactions to be investigated later. This is due to the fact that the VINE decision tree has depth=1 and might end up selecting an explanation that has almost the same significance as another, not informing the end user of the other almost equally-possible explanations. Moreover, being based in ICE plots, it remains prone to the problem of extrapolation.

2.4.3 Local Prediction Methods

2.4.3.1 What-If Plots

Ceteris Paribus Plots map out model predictions around an observation when a single feature is changed, while others remain constant [58]. This might be useful to understand whether the model is locally stable for a certain prediction.

2.4.3.2 Local Surrogate

Ribeiro et al. [2] explain individual predictions by learning an interpretable model locally, using the well-established Local Interpretable Model-Agnostic Explanations (LIME). LIME tests the behaviour of the black-box predictions when the input data is perturbed, drawing from a normal distribution. A new dataset is generated, with weights being attributed with the proximity to the analysed prediction. Using the synthetized dataset, LIME trains an interpretable model, locally accurate to the values predicted by the machine.

Despite being easily interpreted, this method is unstable [59] and not always the most consistent, as changing the neighborhood size or the sampling can radically change the explanations. Recent work by Ribeiro et al. [60] solve the later problem, though only for classifiers. Guidotti et al. [57] suggest the usage of Local Rule-Based Explanations, which learn a decision tree on a genetic generated neighbourhood to provide a local explanation, composed by a logic rule together with a set of counterfactual rules.

Local Interpretable Visual Explanations (LIVE) present a modified implementation of LIME, in which the local exploration data set includes training data in the vicinity, generated by perturbing the explained instance feature by feature, with all the neighbor points being attributed the same weight [61].

2.4.3.3 Prediction Decomposition

SHapley Additive exPlanations (SHAP) offer global consistent and local accurate explanations, ascertained by the game theory Shapley values [62]. For each target value predicted, SHAP values explain the difference between the global average and the obtained prediction, assigning each feature to a contribution for that difference [43, 63, 64]. For the computation, in simple terms, the target variable is calculated for all possible sets (coalitions) of features, with and without the feature of interest. The difference between both is called the marginal contribution and averaged indicates the SHAP value.

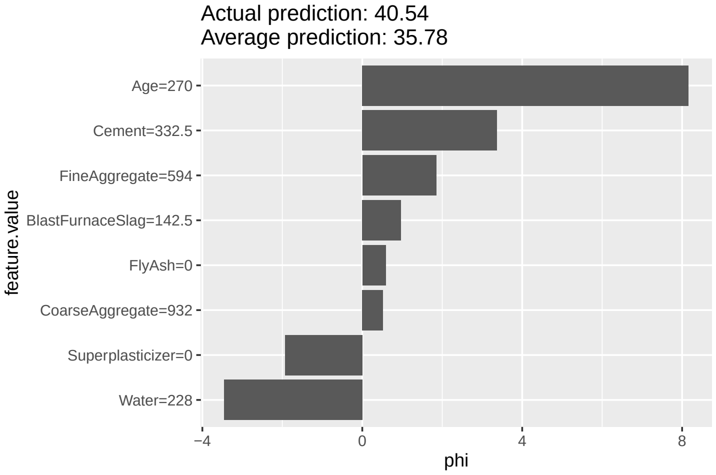


Figure 2.7: SHAP value for a single instance of data set *cementStrength* (c.f.Table 2.1) when trained with a Random Forest (obtained using the R package *iml* [51]). Each horizontal bar represents the contribution of each feature in computing the prediction in comparison to the average predicted value.

However, this method might include unrealistic data combinations when features are correlated. Furthermore, the computation is time-demanding, even though a SHAP time-optimized version for tree ensembles is available [63]. BreakDown [61] presents a fast approximation of the SHAP values, based on model relaxations. Datta et al. [65] presented Quantitative Input Influence (QII), that measure the degree of influence of the features for single or group predictions while accounting for correlated features, however only for classification tasks.

Lundberg and Lee [64] recently presented SHAP values for interaction effects, as well as SHAP dependence plots.

2.4.3.4 Local Feature Importance

Individual Conditional Importance (ICI) and Partial Importance (PI) desegregate the PFI, plotting the expected feature importance in relation to that same predictor variable values [66]. We could also interpret Prediction Decomposition methods as being Local Feature Importance methods, as these inform of the contribution for each feature for a local prediction.

2.4.3.5 Explanations

Unconditional counterfactual explanations indicate examples of small changes in feature values that are expected to alter the predicted result [67]. Nevertheless, this method is affected by the *Rashomon effect*, which affirms that the exact same output can be achieved with contradicting explanations. Prospector is an example of a visual tool that uses PDPs to generate data with different outcomes, by changing the values of the predictors [68].

2.4.4 Summary of Interpretability Methods

Table 2.3: Summary of *Post-Hoc* Interpretability Methods.

Method	**Local/Global**	**Conveyed with**	**Training Data as Input**	**Model Agnostic**	
Regressional RelieF [41]	G	Measure	N	Y	Feature Importance
PFI [14]	G	Measure	Y	N	
MCR [42]	G	Measure	Y	Y	
ICI and PI [66]	G	Visual	Y	Y	
SHAP Feature Importance [64]	G	Visual	Y	Y	
PDP [45]	G	Visual	Y	Y	Feature Effects
ICE [48]	G	Visual	Y	Y	
ALE [50]	G	Visual	Y	Y	
H-Statistic [52]	G	Metric	Y	Y	Feature Interactions
VIN [53]	G	Visual	Y	Y	
VINE [39]	R	Visual	Y	Y	
SHAP interaction effects [63]	L	Visual	Y	Y	
Global Surrogate	G	Model	N	Y	Surrogate
LIME [2]	L	Model	N	Y	
Anchor [60]	L	Model	N	Y	
Live [61]	L	Model	Y	Y	
LORE [57]	L	Explanation	N	Y	
Ceteris Paribus [58]	L	Visual	N	Y	What If
SHAP [64]	L	Visual	Y	Y	Prediction Decomposition
BreakDown [61]	L	Visual	N	Y	
Prospector [68]	L	Visual	Y	Y	Counterfactual Explanations

Chapter 3

Explaining the Performance of a Black-Box

This chapter focuses on the assessment of the performance of black-box regression models in respect to the values of the predictors. We propose three new visual tools, testing them with the data sets and models introduced in Section 2.2.2 to attest their validity and demonstrate their utility.

3.1 Methodology

In light of previous investigation, we concluded that there has been a great development in the field of interpretability methods to the detriment of investigation of accountability methods, which have yet not seen any recent innovation considering regression tasks. Moreover, we found that performance tools that assess regression models solely address the error or the error tolerance in relation to the target value, never establishing a relationship with the values of the predictors. Therefore, our main goal is to design and formulate accountability methods that focus on explaining the reasons for a certain error behaviour by inspecting how changes across the domain of a predictor variable affect the expected model performance.

All the Sub-Sections entitled *Illustrative Examples* (3.2.1.2, 3.2.2.1, 3.2.4.2, 3.3.1.2) provide use cases in which the error behaviour of a black-box regression model was analysed. We have used 18 regression data sets whose properties are described in Table 2.1. For each of these data sets we have estimated the error of the black-box models shown in Table 2.2, using Cross Validation.

Given the number of data sets, predictors and models, we can not show all the resulting plots. For this reason, we only show a few examples that illustrate the power of proposed tools. The full graphs can be seen in the web page `http://github.com/inesareosa/MScThesis`. The same web page contains all code used to obtain the graphs, ensuring full reproducibility of our results and analysis.

All approaches were implemented in R [27] using the *ggplot2* [28] R package.

3.2 Relating the Error with Predictor Values

Determining the expected prediction error for a certain range of values of a feature will provide an estimate of the risk involved in using the model in the domain of interest. Hence, given the feature values of a test case, the end user can decide if the risk is outside the tolerance limits and whether the model is suitable for the task.

In the present section we present two visual tools that graph the functional relationship between the predictor variables values and the expected error of a regression model, suitable for understanding the performance of non-interpretable models for a given test case. These are named Error Dependence Plots (EDPs) and Parallel Error Plots (PEPs) and both are in respect to a single model. The two methods are formally introduced and illustrated with several examples from the benchmark data sets in Section 2.2.2, with the reliability of the EDPs being subsequently evaluated recurring to metric and visual tools.

The proposed tools are based on estimates of the expected error of the models. As such, an important concern relies on the method used to estimate the errors. Using the algorithms on the exact same data used to train the model leads to unreliable and overly optimistic estimates of the error. Thus, we propose to use Cross Validation (CV) to obtain the estimates of the error of the model for each of the available instances. More precisely, a 10-fold CV is used to compute the prediction of the model for each data observation. Using CV, the total data set is split into 10 groups that will serve as test sets later. For each test set fold (hold out), the model is trained with the rest of the data set and evaluated with the respective hold out. This means that each of the instances in the data set belongs to one of the test sets in the folds, and therefore we can compute a prediction using a model that was not trained with that case. Comparing this prediction to the true value we obtain a reliable estimate of the error of the model. Algorithm 2 describes this process in more detail.

Algorithm 2: Obtaining reliable Cross validation Error Estimates.

```
input  : data set $\mathcal{D}$
input  : algorithm $\mathcal{A}$
input  : nr. folds $k$
input  : error metric $Err$
output: error estimates $\hat{E}$
$\mathcal{D}' \leftarrow$ Permute($\mathcal{D}$) // randomly permute the data
$P \leftarrow$ Partition($\mathcal{D}'$, $k$) // create $k$ equal-size partitions
$\hat{E} \leftarrow \{\}$
foreach $p$ in $P$ do
    $M \leftarrow$ Train($\mathcal{A}$, $\mathcal{D}' \setminus \mathcal{D}'_p$) // train $\mathcal{A}$ on all but the partition $p$ cases
    $\hat{e}_p \leftarrow \{$ Err($\hat{y}, y$) $\mid \langle \mathbf{x}, y \rangle \in \mathcal{D}'_p \wedge \hat{y} =$ Predict($M$,$\mathbf{x}$) $\}$ // error of the model predictions
        for the partition $p$ cases
    $\hat{E} \leftarrow \hat{E} \cup \hat{e}_p$
return $\hat{E}$
```

3.2.1 Error Dependence Plots

Error Dependence Plots (EDPs) show the distribution of the estimated error in relation to the values of the predictor variables through the usage of boxplots. These convey the information for a single predictor, but have variants (Bivariate EDPs or Trivariate EDPs) that explore the interaction between two or three predictors.

A boxplot is a graphical tool that depicts the distribution of some data, reporting on the median error as well on quantiles, the upper whisker, the lower whisker and outliers [69]. The first (1Q) and third (3Q) quartil indicates the value in which is located the 25% and 75 % percentil, respectively. The lower whisker is computed with $(Q1 - 1.5(Q3 - Q1))$ while the upper whisker is given by $(Q3 + 1.5(Q3 - Q1))$.

3.2.1.1 Formulation

The core idea of the EDPs is to present the expected error (estimated using Algorithm 2) on the Y-axis, against the values of a predictor variable in the X-axis. However, to calculate the estimated error we require both the value predicted by the model and the true value, which is only known for the training cases. Hence, plotting the error distribution for each possible value of a numeric predictor can be problematic, since each value might not repeat often in the available data, specially when dealing with small data sets. To overcome this adversity, we opted for discretizing the numerical predictor variables into meaningful bins, allowing us to collect several error values for each bin and thus approximate the distribution of the error for these values.

Ideally, these bins should be selected by a domain expert that has a particular interest in evaluating the performance of the model in a set of specific ranges of the variable. For the cases in which this know-how is not available, the bins can be selected using the quantiles of the distribution of the variables. For the cases in which we have no knowledge concerning the domain, as the illustrative examples presented further, we suggest a division of values using the following quantiles: $[0, 10\%]$ (extremely low values), $[10\%, 35\%]$ (low values), $[35\%, 65\%]$ (central values), $[65\%, 90\%]$ (high values) and $[90\%, 100\%]$ (extremely high values).

In case the predictor is nominal, this practice is not necessary since the variables are already discrete. Nevertheless, features constituted by a large number of categories might compromise visualisation. For such situations we suggest either prioritising some of the categories and merging all the remaining or grouping the categories into larger categories, with the help of a domain expert.

Having calculated the bins for all predictors, we can obtain the error of the black-box model for all training cases that belong to each bin. Algorithm 3 describes the process for obtaining the EDP of a predictor variable. This method assumes as input the data set (D), as well as the error estimates $\hat{E}$ for data D (in here obtained using Algorithm 2) and the information on the bins B for a given predictor variable X^k. This algorithm partitions the error estimates $\hat{E}$ according to the respective predictor values, annexing each error $e_i \in \hat{E}$ to the bin b in which the instance x_i^k belongs. Lastly, a boxplot is drawn for each bin of the predictor variable, representing the expected error distribution for that range of values.

Additionally, below each bin we provide information on the number of training cases belonging to

Algorithm 3: Obtaining the EDP of a predictor.

```
input  : data set $\mathcal{D}$
input  : error estimates $\hat{E}$ for the cases in $\mathcal{D}$
input  : bins $B$ of the predictor $X^k$
if $B$ is empty then
    if $X^k$ is numeric then
        $B \leftarrow$ DefineBins($X^k$) // get the bins of $X^k$ using quantiles or end-user ranges
    if $X^k$ is nominal then
        $B \leftarrow$ Categories($X^k$) // get the bins of $X^k$ using categories
foreach $\langle \mathbf{x}_i, y_i \rangle$ in $\mathcal{D}$ do
    $b \leftarrow$ FindBin($x_i^k$, B) // get the bin of the value of $X^k$
    $E_b \leftarrow E_b \cup e_i$ // $e_i \in \hat{E}$ is the estimated error of this case
foreach $b$ in $B$ do
    DrawBoxPlot($E_b$)
```

each bin and the respective percentage of the full data set. For comparison, EDPs show the error distribution over the entire data set on the right side of the plot as well as a dashed line through all the plot indicating the value of the median expected error.

3.2.1.2 Illustrative Examples

Figure 3.1 shows two examples of EDPs for the data set *a1* (Table 2.1). The leftmost plot shows an example for a nominal predictor - *size*, in this case with 3 possible values shown in the X-axis of the plot $\{$*small*, *medium*, *large*$\}$. For each of these values we see the corresponding expected error distribution of a Gradient Boosting Machine (GBM) (with parameters in Table 2.2). From this simple example we can observe that this model performance varies considerably depending on the value of *size*. More specifically, the GBM is expected to have worse performance on test cases where the value of the predictor is *small*. When the value is *medium* the performance seems considerably better. In the right side of the EDP, it is shown the overall performance of the model, i.e. without any conditioning on the value of the predictor under analysis.

The rightmost plot of the figure shows another example for the same data set, this time for the *Cl* numeric predictor, which was discretized into several bins according to the quantiles, as described above. Here we observe that the GBM performs considerably worse on lower values of *Cl*, while for extremely high values of this variable the performance is much better.

In some situations some of the bins of a predictor may have an estimated error that can be regarded as an outlier when compared to the most common errors of the model (here represented as a small dot). This may distort the Y-axis scale of the EDP hiding some of the information that allows the effective comparison between the bins. In these cases we recommend the use of the logarithmic scale in the Y-axis. Such situation is illustrated in Figure 3.2. As you can see with the normal scale (left plot) the presence of some outlying errors makes the comparison among the bins strenuous. The right side graph shows the same error distributions using a log scale, where we can better observe some considerable

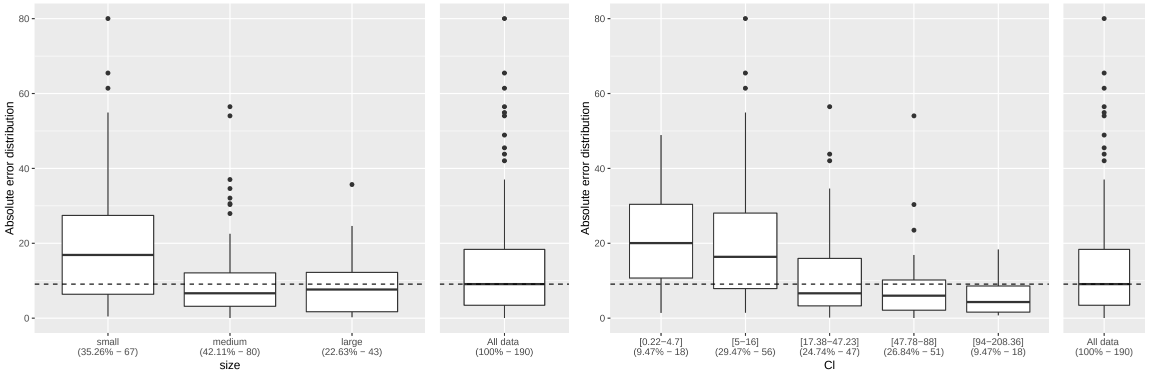


Figure 3.1: Error Dependence Plot for data set *a1* trained with GBM for features *size* (Left) and *CI* (Right).

differences between the bins.

In here we use a logarithmic scale calculated with $\hat{E} = log(1 + |\hat{y} - y|)$ to always guarantee a positive scale in the plots. This alternative was fit for the perspective of a mathematical illiterate end user, which could easily fall into the pitfall of assuming that a negative $\hat{E}$ in the plot would signify a negative error.

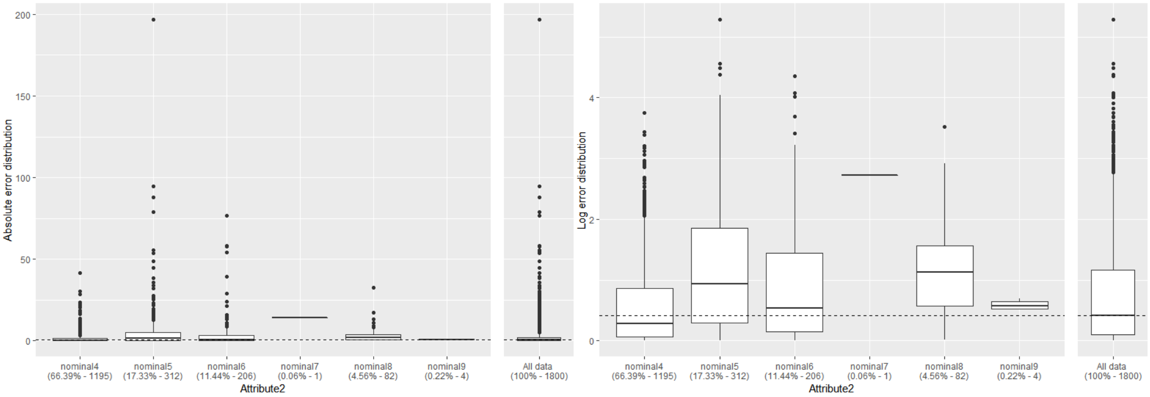


Figure 3.2: EDP for dataset *maxTorque* trained with RF for feature *Attribute2*. Left: Absolute Error, Right: Log Error

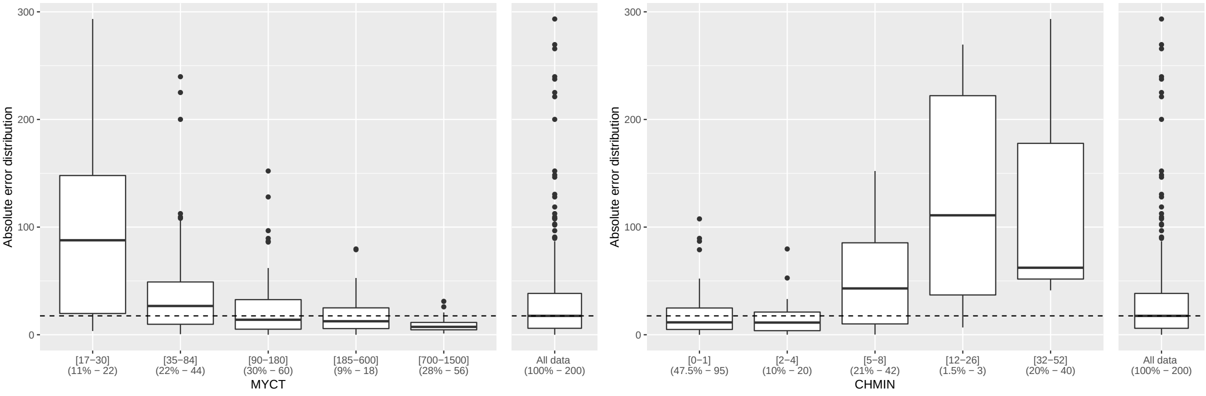


Figure 3.3: EDP of a GBM for features *MYCT* (Left) and *CHMIN* (Right) of data set *machineCpu*.

Figure 3.3 shows two EDPs of a GBM for two numerical features of data set *machineCpu*: *MYCT* (on the left) and *CHMIN* (on the right). For *MYCT*, one can observe that the performance improves with higher values of this feature, as lower values of the predictor variable show higher expected median error than globally, whereas in high and extremely high values (*MYCT* = $[185, 600]$ and *MYCT* = $[700, 1500]$) the opposite occurs, with the boxplots indicating lower expected errors. The plot on the right dissects the model performance for the range of *CHMIN*, where we observe that the GBM underperforms when this predictor has central or high values (3 rightest bins). In fact, these bins help explaining the higher estimated errors presented as outliers in the boxplot for the overall data. For low values (*CHMIN* = $[0, 1]$, *CHMIN* = $[2, 4]$), the model has a far better performance. This EDP provides clear indications that the GBM is not as reliable for high values of *CHMIN*, which may be a very useful information for the end user. In both cases, some of the bins show high variability (*MYCT* = $[17, 30]$ *or CHMIN* = $[32, 52]$) in shape of a wide boxplot. This somehow may serve as an alert that there might be other factors influencing the behaviour of the model in those ranges of values. The larger this error variability (particularly the box that is supposed to contain 50% of the values), the more careful the end user should be in drawing conclusions from the individual boxplots.

3.2.2 Bivariate EDPs

The interactions between predictors may have an impact on the performance of the models. However, these are ignored by regular EDPs, thus motivating the development of a new tool that provides insight into possible interactions - the bivariate EDPs. These graphs are conceptually the same as the EDPs described in the previous subsection, but they show the error distribution for a combination of two predictors (X^{k_1} and X^{k_2}). These are obtained with the procedure described in Algorithm 4, which partitions the errors across all possible combinations of bins between both predictors.

Algorithm 4: Obtaining the Bivariate EDP for two predictors (p=1 and p=2).

```
input  : data set $\mathcal{D}$
input  : error estimates $\hat{E}$ for the cases in $\mathcal{D}$
input  : bins $B^{k_1}$ and $B^{k_2}$ of the predictors $X^{k_1}$ and $X^{k_2}$
foreach $p$ in {1,2} do
    if $B^{k_p}$ is empty then
        $B^{k_p} \leftarrow$ DefaultBins($X^{k_p}$) // get the bins of $X^{k_p}$ using quantiles or categories
foreach $\langle \mathbf{x}_i, y_i \rangle$ in $\mathcal{D}$ do
    $b_{\{1,2\}} \leftarrow$ FindBin($X_i^{k_1}$, $B^{k_1}$, $X_i^{k_2}$, $B^{k_2}$) // get the bin respective to the combination
        of values $X^{k_1}$ and $X^{k_2}$
    $E_{b_{\{1,2\}}} \leftarrow E_{b_{\{1,2\}}} \cup e_i$ // $e_i \in \hat{E}$ is the estimated error of this case
foreach $b_2$ in $B^{k_2}$ do
    foreach $b_1$ in $B^{k_1}$ do
        DrawBoxPlot($E_{b_{\{1,2\}}}$)
```

The Bivariate EDP depicts the error distribution in a different subplot for each bin in X^{k_2}, with the bins of X^{k_1} values as the X-axis. Note that if a combination of values of the two variables does not occur in the data, the respective boxplot is not shown. Furthermore, it is also included a comparison subplot, which whether shows the total estimated error distribution or the error distribution when conditioning one of the predictor variables in analysis.

For a deeper insight, EDPs can also be adjusted to capture trivariate interactions, which can be displayed in two different modes: *grid* or *condensed*. The first presents the possible bin combinations as a grid, with the bins for the first predictor (X^{k_1}) in the X-axis, the rows as bins in X^{k_2} and the columns as the bins in X^{k_3}. On the other hand, in the *condensed* version the EDPs are juxtaposed, conditioning the values of X^{k_2} and X^{k_3} and having X^{k_1} values in the X-axis. In this case, if a combination of values of the two/three variables does not occur in the data, the respective subplot/boxplot is not shown. As for the bivariate EDPs, this variant also includes a comparison plot. Two illustrative examples of the two different modes are plotted in Figures A.1 and A.2 of Appendix A

However, we warn that Trivariate EDPs present a more onerous visualization, as the number of bins and subplots escalates. Moreover, these can find their reliability compromised, particularly for small data sets, where some of the bins might not be representative due to their size.

3.2.2.1 Illustrative Examples

Figure 3.4 shows an example of a Bivariate EDP for the data set *a7* when trained with a Support Vector Machine (SVM). In this case we explore the impact of the predictors *season* and *PO4* on the estimated error of the model. The top left graph shows the overall error distribution of the model without any conditioning of the two predictors. The remaining panels show the error distribution across the different bins of *PO4*. For each bin of *PO4*, we show the boxplots of the bins of *season*. This small example allows us to observe that the SVM has a considerably different behaviour for *season = autumn* when *PO4* is in the range $[169, 285.71]$ (high values of *PO4*). We can also observe that for the lowest values of *PO4* the performance is generally much better, independently of the season.

As stated before, if a combination of variables is absent from the data set, the respective boxplot is not drawn, as it is the case of the bottom right panel of Figure 3.4 where there is no boxplot for *season=autumn*. The percentage and absolute numbers of the bins are different from the ones that appear in the respective EDPs, since it is calculated by the size of each combination boxplot in respect to the full data set. For instance, from the top rightmost panel we can infer that there are 9 cases with $season = spring \wedge PO4 \in [13.6 - 69.93]$, which correspond to 4.74% of the data set size.

Figure 3.5 shows the bivariate EDP that corresponds to the two variables with EDPs in Figure 3.3: a GBM for data set *machineCPU*, conditioning the numerical features *MYCT* and *CHMIN*. The top left part of the graph shows the EDP for feature *CHMIN*. As seen before, in Figure 3.3, the GBM underperforms for cases in which *CHMIN* has higher values and when *MYCT* has lower values. This plot brings more insight, informing on the interactions between the two features, which are particularly evident when both conditions occur simultaneously, as these are the values where the boxplots differ the most. In the two subplots corresponding to the lower values of *MYCT* (*MYCT*=$[17, 30]$ and *MYCT*=$[35, 84]$), it is noticeable that the model only has worse performance for the cases in which *CHMIN* has higher values. Relying on this tool, the end user can conclude that the GBM might not be trustworthy when operating in conditions with simultaneously low *MYCT* and high *CHMIN*.

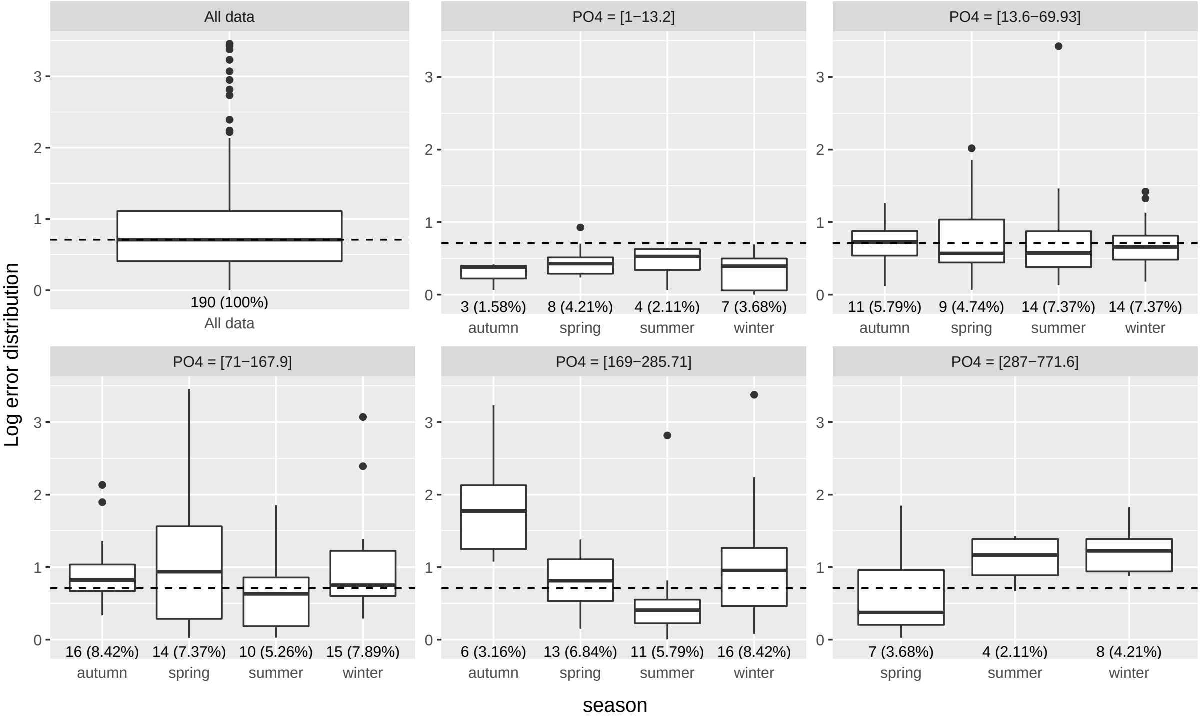


Figure 3.4: Bivariate EDP for data set *a7* trained with SVM for features *season* and *PO4*.

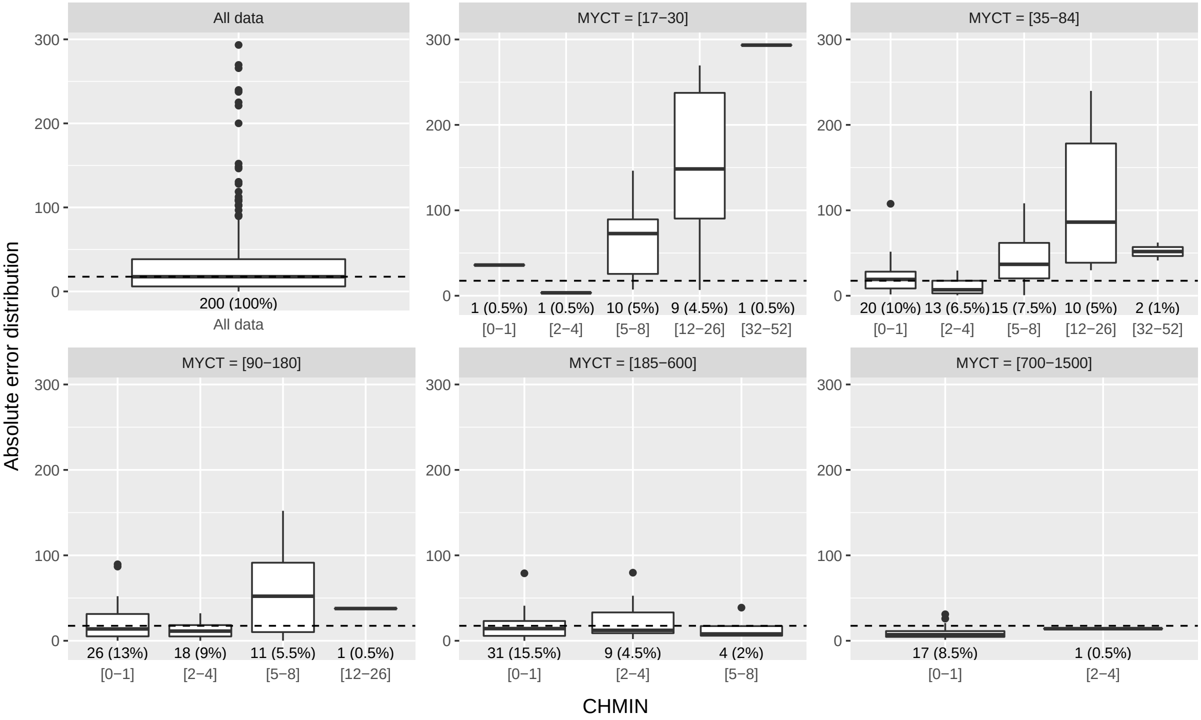


Figure 3.5: Bivariate EDP of a GBM for data set *machineCPU* for interactions between *MYCT* and *CHMIN*.

3.2.3 Evaluating EDPs

In this subsection we address how reliable are the estimates of the error distribution shown through conditioned boxplots. Although these boxplots are obtained with error values estimated through Cross Validation, it is important to evaluate how effective are these boxplots in anticipating what will be the error of the model in future test cases. This is more important given the fact that EDPs are univariate and thus underestimate the impact the interactions among different variables may have on the error.

In order to assess reliability of the distributions shown through the boxplots with the EDPs we have carried out a visual and a metric evaluation.

For both experiments, each data set in Table 2.1 was randomly partitioned into a training (70%) and a test set (30%). The training sets were used to train the models in Table 2.2 and the estimated errors for each instance in the training sets were computed using Cross Validation applied only on the training set. On a second stage, we obtained the predictions on the separate 30% test set that was left out, calculating the respective errors.

For the visual evaluation, the training error estimates (obtained through CV) were used to obtain the EDPs for all predictor variables, with the binning of the predictor values made using the entire data set values, assuming the partitioning was based on a previous knowledge of the domain. Our goal is to check if the distribution of the errors obtained on the test set is similar to the distribution shown on the EDPs obtained using the CV estimates on the training data. If that is the case then we confirm that EDPs obtained using the procedure described in Section 3.2.1 can be used as tools for reliably anticipating the true error of the models in future instances of the same problem.

To facilitate the visual comparison between the distribution of the errors observed on the test set and the distribution shown on the EDPs, we propose a small variant of these graphs in which two boxplots are show for each value of the predictor: (i) the original boxplot of the EDP corresponding to the training set; and (ii) the boxplot of the errors of the model on the test set. Through visual inspection, the end user can verify if the actual error had a similar distribution as the one indicated by the EDP.

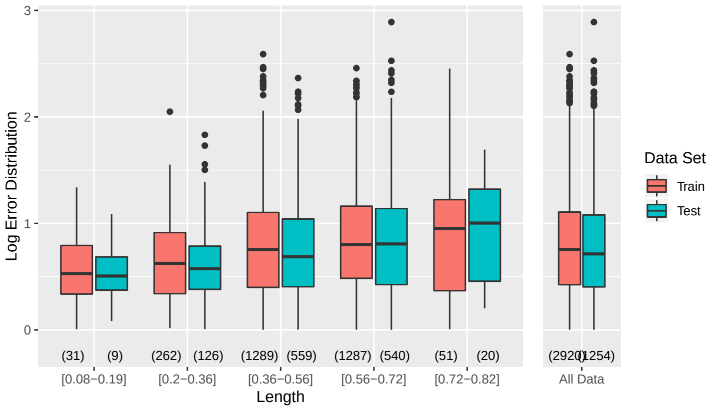


Figure 3.6: Graphic evaluation of EDP for feature *Length* of data set *Abalone* trained with NN.

Figure 3.6 exemplifies the EDP evaluation of a Neural Network (NN) when trained on data set *Abalone*, for feature *Length*. The right part of the plot shows the logarithmic error distribution for the overall data, while the left part dissects the numeric values of *Length* into the usual bins. For each bin we show two boxplots with the leftmost, in pink colour, being the boxplot as obtained by a standard EDP when calculated with the training data, while the rightmost boxplot, in blue colour, represents the actual error distribution observed on a separate test set. Analysing the estimated error distribution using the training set, it can be concluded that the model is expected to have a slightly better performance for low values of *Length* (corresponding to the two left bins), and a worse performance for high values of *Length* (corresponding to the three right bins). As we can observe the test set error distribution follows the same general trend, hence showcasing the reliability of the EDPs information in this particular example. In what regards the variability within each bin, this follows the expected distribution, as it is observed from the similar box size of the boxplots (that represent 50% of the data).

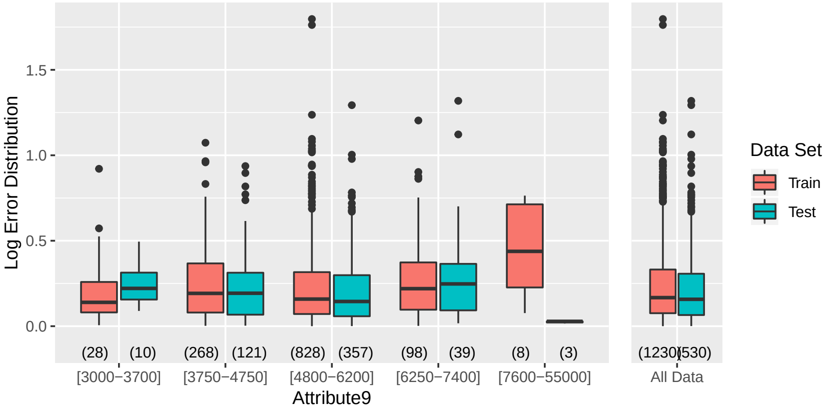


Figure 3.7: Graphic evaluation of EDP for feature *Attribute9* of data set *fuelCons* trained with GBM.

From the analysis of similar experiments carried out with all the data sets in Table 2.1 we can conclude that, as expected, EDPs tend to have higher reliability for larger data sets and, in most cases, for bins with considerable representation. This is caused by the fact that Cross Validation estimates will be more effective when the available data samples are sufficiently large. In Figure 3.7 we show the EDP evaluation of a GBM trained on data set *fuelCons* for feature *Attribute9*. The major disparity between the error distributions is observed for *Attribute9* = [7600-55000], where the number of cases with this value in the training set was of only 8 (0.65% of the training set), and in the test set only 3 cases had this value of the feature. Another bin, with the range *Attribute4*=[3000-3700], presents a smaller discrepancy, having 28 (2.27%) cases on the training set. The other bins, with larger representation, present the best results, with the test error showing a similar distribution to the one predicted by the EDP.

To further evaluate the accuracy of the estimates provided by EDPs, we ran some formal tests that compare two continuous, one-dimensional probability distributions: the Kolmogorov-Smirnov [70, 71] and the Anderson-Darling [72] tests. For both we used the same results and distributions obtained from the division of the data set into a training and test set, as defined above. The two tests assume a null hypothesis that states that the two samples are drawn from the same distribution. In our evaluation, rejecting the null hypothesis would then mean that the EDPs did not estimated successfully the error behaviour for that given range of predictor values.

Kolmogorov-Smirnov can compare two samples (two-sample K-S test), by quantifying the distance between the empirical distribution function (as defined in Equation 3.1) of each sample, being sensitive both to location and shape of the distribution.

$$F_n(x) = \frac{\textit{number of elements in the sample} \leq x}{n} \tag{3.1}$$

The distance between the distribution functions is then defined by:

$$D_{n,m} = \sup_x |F_{1,n}(x) - F_{2,m}(x)|, \tag{3.2}$$

where n and m are the sizes of the samples. Using widely available tables (e.g. [73]), a critical value $D_{n,m,\alpha}$ is calculated according to the chosen confidence level α, with the null hypotheses being rejected at level α if the $D_{n,m} > D_{n,m,\alpha}$.

The Anderson-Darling (AD) Test is similar to the KS test, but is more sensitive at the tails, presenting better results in the occurrence of outliers. Considering

$$H(x) = \frac{nF_{1,n}(x) + mF_{2,m}}{m+n}, \tag{3.3}$$

the distance is calculated with:

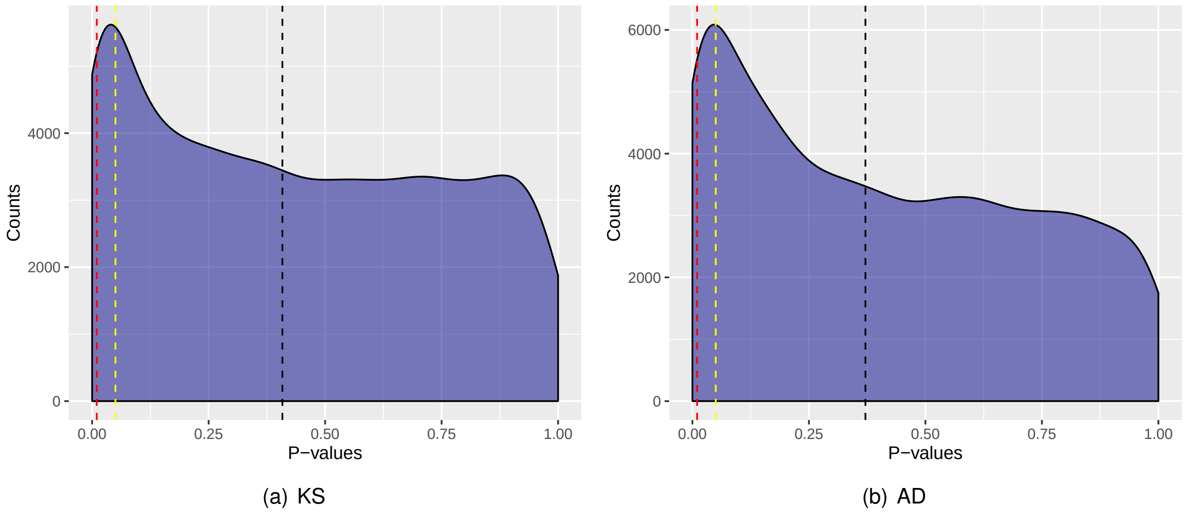
$$A^2_{n_m} = \frac{nm}{n+m} \int_{-\infty}^{\infty} \frac{\{F_{1,n}(x) - F_{2,m}(x)\}^2}{H(x)\{1 - H(x)\}} dH(x). \tag{3.4}$$

For the case of EDPs, this particularity is significant since these outliers might be a product of interactions between predictors, and therefore should not be ignored. Furthermore, this test requires less data than the KS test to reach sufficient statistical power [74].

The two tests were applied using the functions *ks.test* and *ad.test*[1] from R package *kSamples* [75], respectively for the KS and the AD test (as detailed in [76]). In both cases a p-value is returned, which should be compared to a significance level α - if less or equal than α, the null hypothesis should be rejected. In the absence of a predefined significance value, the p-values were analysed on their own and in comparison to two commonly used ($\alpha = 0.01$ and $\alpha = 0.05$) significance values. High p-values indicate there is not a statistical support for assuming a difference between two distributions, so these are the values we are aiming for. All the bins for each feature from all the 18 data sets trained with the 4 different models were analysed, comparing the estimated error distribution sample from the training set to the error distribution sample from the test set.

Figure 3.8 shows the overall density function for the p-values, calculated with KS test and AD test, showing the number of counts for each p-value returned as well as three vertical lines illustrating two of the most commonly used significance levels (α=0.01 in red and α=0.05 in yellow) and the median p-value. We can observe that both tests present an high density of low p-values, while the median presents a satisfactory p-value. KS test presents a more constant density function, while the AD shows

[1] two versions are available, but in here only the second one is used since it accounts for tied observations



(a) KS

(b) AD

Figure 3.8: Density function of p-values. Yellow dashed line indicates α=0.05; red dashed line indicates α=0.01; black dashed line indicates median p-value.

an higher prevalence for lower p-values. Note that the percentage of p-values with values under 0.01 is in fact low, around 8% for both tests, while under 0.05 we are dealing with values around 15%.

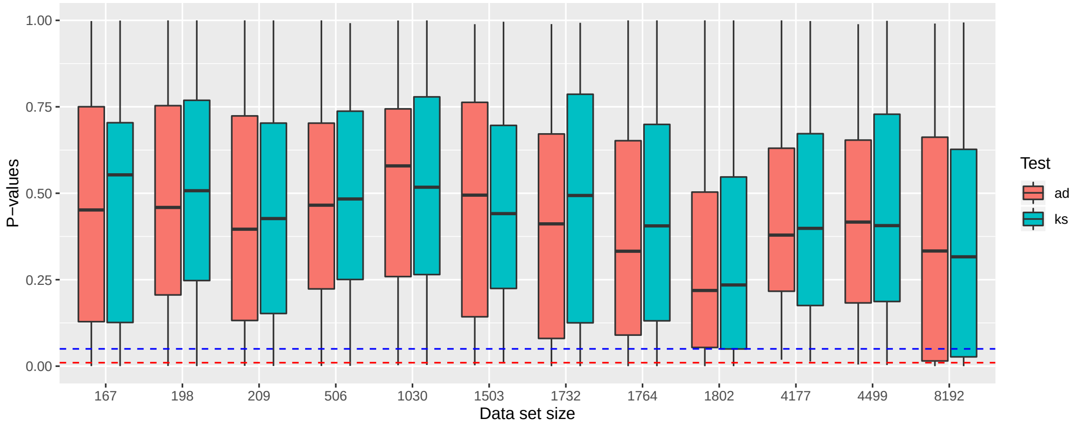


Figure 3.9: Boxplots of the p-values obtained for the two distribution tests (KS and AD) in terms of the size of the data set. *Blue dashed line*: α=0.05, *Red dashed line*: α=0.01 .

Figure 3.9 shows the distribution of the p-values for each test through the usage of boxplots, according to the size of the data set in analysis. We can observe that the values returned by the KS test seem to have a slightly higher tendency comparing to the AD test. We can also notice that for all the cases there are p-values approximately from 0 to 1, what shows the variability within the values returned for each data set and model. In dashed lines are represented two significance levels: α=0.01 in red and α=0.05 in blue. All the boxes of the boxplots, which represent the middle 50% of the distribution, are situated above the 1% significance level, but the same cannot be stated if α is chosen to be of 5%, as the data set with size 8192 includes p-values ≤ 0.05 in the middle box, indicating a relative high probability of the EDP providing a wrong estimate of the error distribution for this particular case. Apart from that situation, we can state that the values found are adequate, with most data sets having at least 75% of the metric with good values, with both tests. We observe that for data sets up to 1030 instances, there appears to be a slight tendency of an increase of the p-values with the number of instances in the data set (except for the case with size=209), corroborating our findings in the visual evaluation. However, it is hard to establish a pattern for larger data sets, which present variable results with the size of the data set. This variability might arise from particularities of the data sets and models, but it should be taken into consideration that the calculation of both tests accounts for the sample size, and therefore this might affect any correlation that might exist between the efficacy of EDPs and the size of the data set.

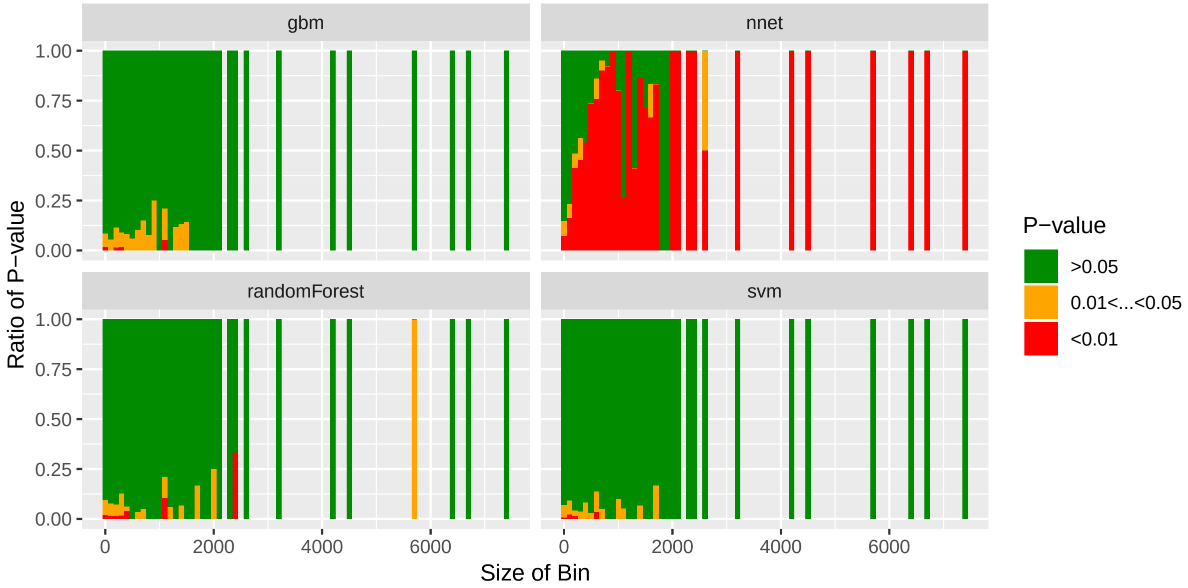


Figure 3.10: Proportion of p-values against size of bin for each model (GBM, NN, RF and SVM), distinguished if above or below $\alpha = 0.05$ and $\alpha = 0.01$. Bar width of 100 instances.

Figure 3.10 analyses in more depth the AD test p-values against the significance levels $\alpha = \{0.01, 0.05\}$ for the size of the predictor bins of each data set, for each algorithm. Here we distinguish between the percentage of p-values above $\alpha = 0.05$ (in green), below $\alpha = 0.05$ (in yellow and red) and below $\alpha = 0.01$ (in red), plotting the ratio of p-values for each size of the bin. We observe a particularity for the Neural Networks, which the CV estimates clearly have the worst performance in predicting the error behaviour, as we can conclude by the higher percentage of low p-values. All the other models, specially the GBM and the SVM, show an improvement on the reliability of the estimates with the increase of the size of the bin, strengthening the assumptions made in the previously discussed visual evaluation.

In summary, our visual and metric analysis show that EDPs obtained with errors of a black-box mode estimated through a Cross Validation process will generally provide reliable feedback concerning the expected error of the model for each feature value, if enough data is available for this Cross Validation process. However, a special advert should be made to the usage of EDPs with Neural Networks, since the respective estimates were found to not consistently provide trustworthy results. These results do not arise from a failure of the EDPs but rather from the fact that the estimated CV errors have been observed to be unreliable, due to an higher instability of the NNs that lead to the performance on a training set not being completely replicable on a test set.

3.2.4 Parallel Error Plots

EDPs have limitations when plotting multiple variables simultaneously. In fact, for neatly achieving results these are restricted to display the information for a maximum of three variables at a time, ignoring interactions between a larger number of features. However, most real world problems tend to have many more predictor variables. In this context, we propose a new multivariate representation of the performance of a black-box regression model that complements the univariate perspective of EDPs: the Parallel Error Plot (PEP).

PEPs have the same high-level objective as EDPs, of informing the end user on the dependency between the expected error of a black-box model and the predictors. But instead of showing this dependency with boxplots over the values of a single predictor, PEPs represent this error profile across the range of values of numerous predictors at the same time.

3.2.4.1 Formulation

For the elaboration of PEPs we propose using parallel coordinate plots [77] as a visual tool for informing the user on the error profile of a black-box regression model across a set of predictor variables.

Given the limitations of parallel coordinate plots, a boxplot of the error distribution could not be shown for each bin of all the predictors. In order to mitigate this, we decided to split the very high errors from the rest. The assumption is that end users are particularly interested in knowing the conditions that lead the models to a unusually bad performance as these are indicators of higher risk. In this context, we suggest dividing the error profile in two categories: (i) the top 10% higher errors; and (ii) the rest of the errors. The concept of PEPs is not limited to this specific division and the user is free to select other splitting criterion, particularly when dealing with extremely large data sets, where the percentage should be adjusted to avoid jeopardizing visualization.

Parallel Error Plots use a method analogous to parallel coordinate plots, in which each variable is shown on the X-axis and is represented by a vertical bar that results from standardising the scale of each variable with the goal of having them represented in the same Y-axis coordinate. This uniformisation process can be achieved through several ways, being that PEPs use a method that maps the original range of each variable into a [0,1] scale, with 0 corresponding to the minimum and 1 to the maximum values of the variable in the data set. Mapping all feature values to this uniform scale supports the display of all values on the same Y-axis.

Using this scale, each observation in a data set is then represented by a line that crosses each vertical bar (representing a feature) according to the respective scaled value of the variable. On top of this, PEPs color the line of each case according to the error of the black-box model it is trying to explain. Specifically, if the model had a very high expected error (by default on the top 10% largest errors) when forecasting some case the corresponding line is shown in red, otherwise the line is drawn in grey. This enables the end user to easily detect some patterns and overall tendencies concerning the conditions in terms of the predictors that lead to higher prediction errors. For data sets with more than 1000 data instances, each line is plotted with a certain transparency level to facilitate the identification of patterns

- if several data instances have similar behaviour, the lines will overlap and consequently become more visible.

We advise ordering the predictors in the X-axis by a score of feature relevance, using feature importance methods such as the ones proposed by Breiman [14] or Fisher [42]. This way end users can easily confirm whether the lowest performance is explained by the most important features. For the benchmark examples provided in this section, the predictors were ordered by importance using the function *varImp* from the R package *caret* [78], which calculates the variable importance through model-specific methods.

Our implementation of PEPs also holds the option of a more detailed representation of the top errors. This consists of representing the value of the expected error using a level of saturation of the red color, instead of only separating the high errors. In order to avoid misjudgment, the user should have into consideration that this alternative strongly enhances outlier error instances if their values fall far from the rest of the other errors.

PEPs present some limitations when outliers occur in the predictors range, since this leads to a compression of the other values. This limitation can be addressed by using other methods of making the scales of the variables uniform, that are robust to outliers. Furthermore, with a large data set, the visualisation might get confusing. In these cases, one can either randomly subset the data set to help interpretation or pick a lower percentage for defining the top errors. Finally, the visualization provided by PEPs may also suffer from an excessive number of predictors. In these cases we can opt for showing only a subset of the predictors, which could be determined by the scores of feature relevance introduced beforehand.

3.2.4.2 Illustrative Examples

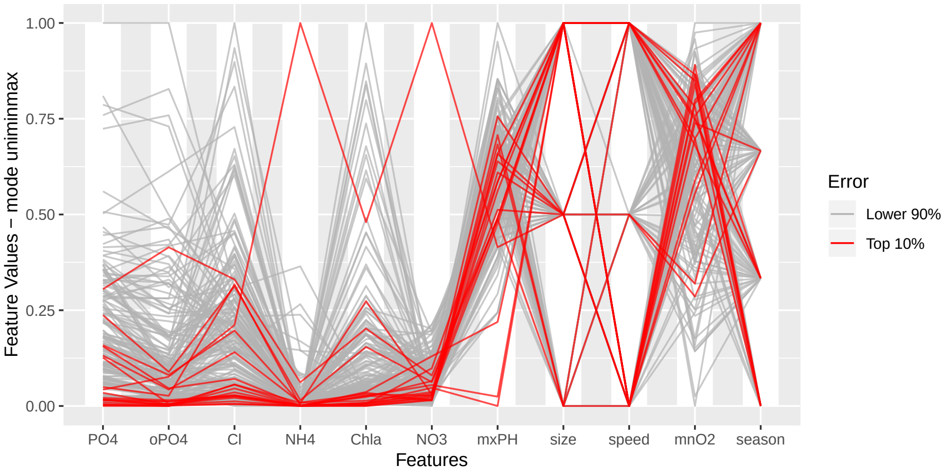


Figure 3.11: Parallel Error Plot of Random Forest for dataset *a1*.

Figure 3.11 shows an example of the Parallel Error Plot (PEP) of a Random Forest (RF) trained on data set *a1*, with all predictors of the data set ({*PO4, oPO4, Cl, NH4, Chla, NO3, mxPH, size, speed, mnO2, season*}) ordered from left to right in increasing order of estimated feature importance. PEPs help in identifying interesting patterns concerning the largest errors of the models. In fact, with this plot it can be observed that the largest errors of the Random Forest occur for the cases in which *PO4, oPO4, Cl, NH4, Chla* and *NO3* have lower values of their range and when *mnO2* has higher values . On the other hand, highest errors do not seem to be strongly correlated with the values of the variables *season*, *size* and *speed*. This type of information can be of great value when deciding whether we can trust the prediction of the Random Forest for a new test case of this problem.

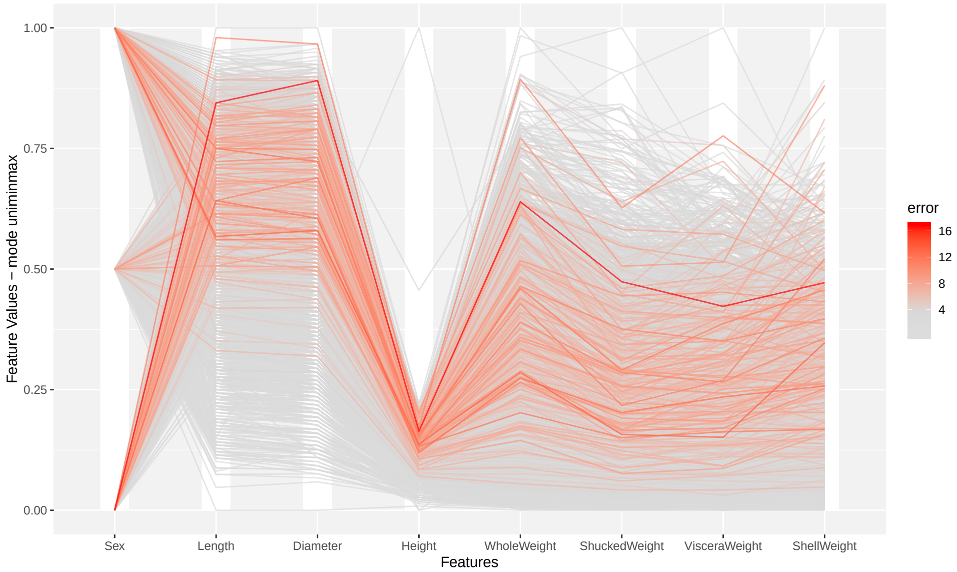


Figure 3.12: Parallel Error Plot for data set *Abalone* trained with GBM with increasing saturation with error value.

Figure 3.12 depicts the alternative representation for data set *Abalone* when trained with a GBM. As in the example before, the predictors in the X-axis are ordered according to the importance score. The difference relies in the top 10% estimated absolute errors, coloured in red, in which the higher errors are represented with increasing saturation. Inspecting through all feature variables, is noticeable the prevalence of high errors for high values of *Length* and *Diameter* as well as for low values of *Height*. Features *WholeWeight*, *ShuckedWeight*, *VisceraWeight*, *ShellWeight* do not present any high error for extremely low values within their range. Moreover, the categories in *Sex* do not appear to have a strong correlation with the performance of the model.

3.3 Comparing the Performance of Regression Models

EDPs and PEPs allow to understand the conditions in terms of predictor values that lead to different predictive performance of the models. Both methods add an explanation of these differences, as a function of the predictor values. In this section we argue that it is also relevant to do this analysis with the goal of comparing different models on the same problem. This would allow to make model selection for specific test cases based on their predictor values. With this goal we propose the Multiple model Error Dependence Plots, in their univariate and bivariate forms.

3.3.1 Multiple Model Error Dependence Plots

Multiple model Error Dependence Plots (MEDPs) extend the idea of EDPs to further analyse multiple models simultaneously, across the range of a predictor. The tool was tested plotting for each predictor of the 18 data sets from Table 2.1, comparing the four models defined in Table 2.2.

3.3.1.1 Formulation

Similarly to EDPs, we start by compartmentalizing the feature of interest using the process described in Section 3.2.1. The errors of the data set are then obtained for each model using Cross Validation (Algorithm 2). Finally, for each bin within the predictor values, the error boxplots of each model representing the estimated error distribution are arranged side by side, enabling the end user to scrutinise and compare the performance of the models across the domain of the predictor.

In order to facilitate comparisons, the right section of the MEDPs shows the overall performance of the models. Furthermore, a dashed line crosses horizontally the entire plot representing the median predicted error for each model.

Through visual inspection of the estimated error behaviour, this tool helps finding the model most suitable for any particular test case given its feature values.

3.3.1.2 Illustrative Examples

Figure 3.13 depicts the MEDP of the four models (SVM, RF, NN and GBM) for the bins of feature *BlastFurnaceSlag* of the data set *concreteStrength*, where the models present a similar estimated error behaviour across all values of the variable domain. In such cases, the choice of model based on the performance seems independent of the predictor variable values under study (*BlastFurnaceSlag* in this example).

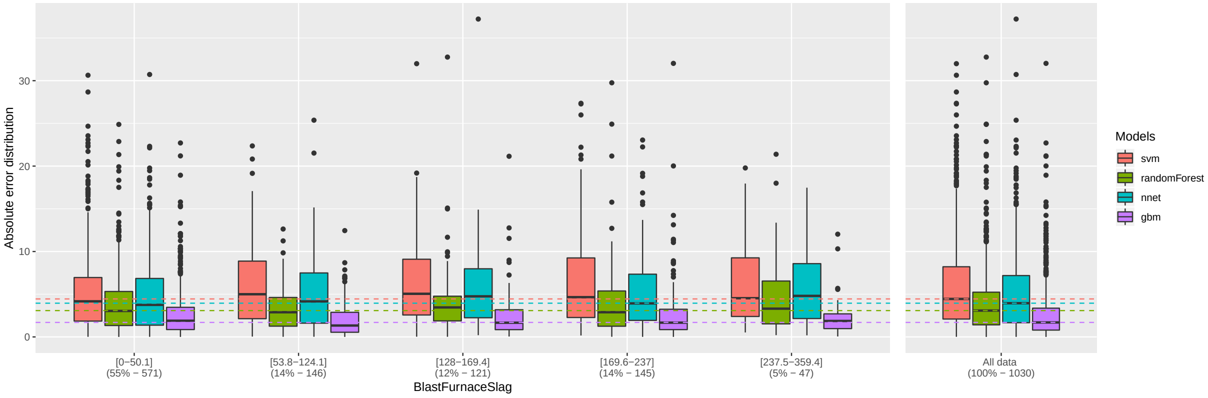


Figure 3.13: MEDP for feature *BlastFurnaceSlag* in data set *concreteStrength*.

MEDPs show particular competency in identify whether the model with the best global performance is outperformed for a certain range or category of predictor variables. As an illustrative example, take the MEDP on the *acceleration* data set for *Attribute1* (Figure 3.14), in which the best overall performance is achieved with the Gradient Boosting Machine, as seen in the right part of the plot. The four dashed lines represent the median predicted error for each of the models. Analysing the performance of all 4 models for each category of *Attribute1*, one can conclude by the expected error distribution (conveyed with an higher median absolute estimated error and with the position of the boxplot indicating higher estimated errors) that this model in fact underperforms when the feature is *Attribute1=nominal5*. This plot then indicates that, if analyzing cases where *Attribute1=nominal5*, the Gradient Boosting Machine is not as reliable as the other models, in contrary to the what the overall performance illustrates. Note that the bin *Attribute1* = *nominal4* does not have the representation necessary to reach a reliable conclusion in respect to the performance of the models.

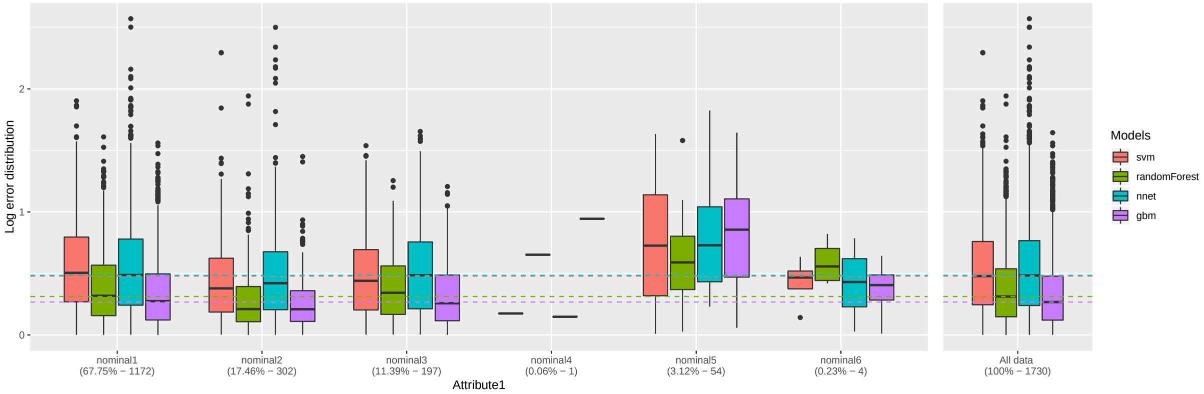


Figure 3.14: MEDP for feature *Attribute1* of data set *acceleration* using logarithmic scale.

MEDPs can also be a useful tool to decide between two models with a similar overall performance. Figure 3.15 shows the MEDP of data set *cpuSm*, for the numerical feature *runqsz*. In the rightest part of the plot, which depicts the overall estimated performance of the models, the performance of the Random Forest and the Gradient Boosting Machine appear to follow the same tendency, with identical median error, boxplots and outlier values. However, drilling down the range of *runqsz*, the user can perceive that the Random Forest outperforms the Gradient Boosting Machine for central and higher values of this feature (bins of range $[184, 430]$, $[434, 874]$, $[896, 2823]$). Although these are bins with little expression in terms of percentage of the data set (together these account for 2.6%), this can be a crucial information when opting for a model, particularly if operating in these ranges of values.

3.3.2 Bivariate Multiple Model Error Dependence Plots

As discussed before, in Section 3.2.2, the analysis of a single variable hides intricate relations with other variables, which might affect the performance of the models.

Hence, we suggest extending the analysis of the comparative performance of several models to

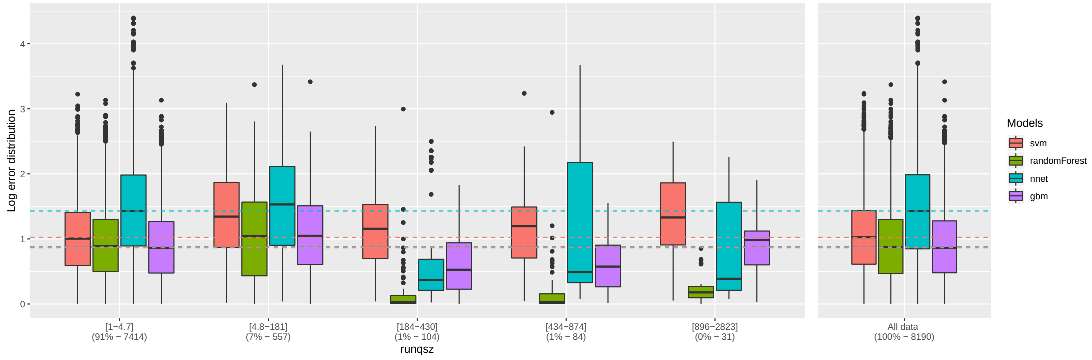


Figure 3.15: MEDP for feature *runqsz* of data set *cpuSm* using logarithmic scale.

two variables simultaneously, allowing for the identification of interactions between predictors. Bivariate MEDPs visualise the estimated error distribution for all possible combinations of bins of two predictors. In similarity to Bivariate EDPs, this method fixes the bins of one of the variables on the X-axis and then plots, through different facets, the MEDPs across the values of the second feature, thus having each facet representing the conditioned estimated error distribution. In order to enable comparisons, a plot is included with either the MEDP for one of the predictors or the error distribution for the overall data.

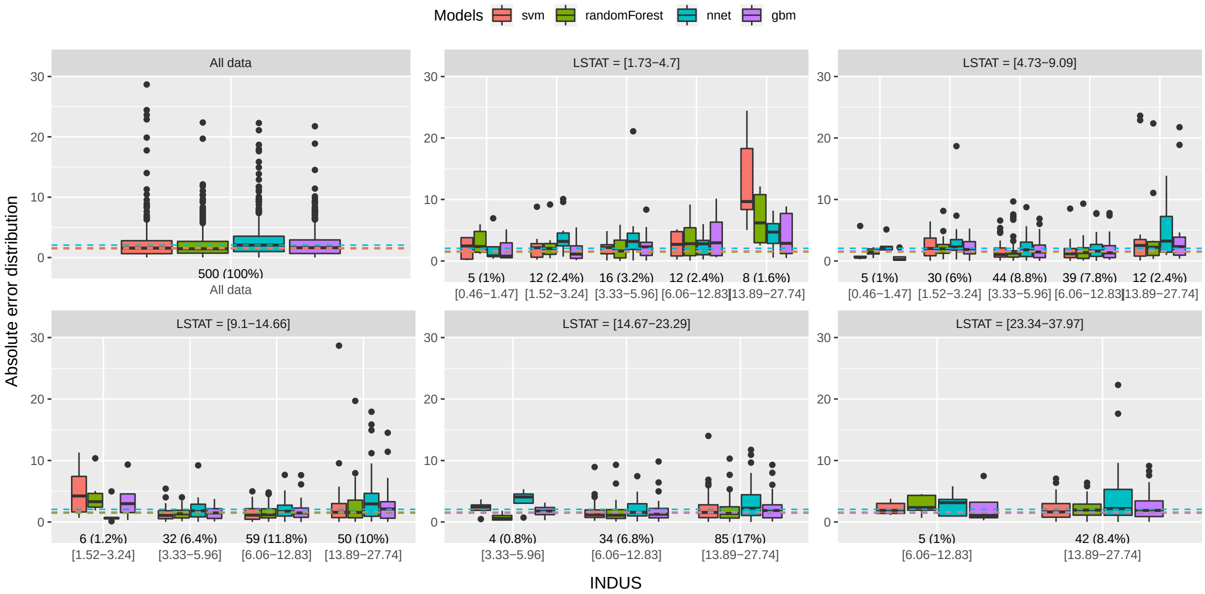


Figure 3.16: Bivariate Multiple model Error Dependence Plot for features *INDUS* and *LSTAT* of data set *boston*.

One example of a Bivariate MEDP is pictured in Figure 3.16, which shows the plot of data set *boston* for numeric features *LSTAT* and *INDUS*. Each of the bins of *LSTAT* serves as a conditioning - represented as a facet (subplot). For each of these facets we have the MEDP across the values of *INDUS*, conditioned on the respective value of *LSTAT*. Note that when a combination of bins is not present in the data set, it does not appear in the Bivariate MEDP, as for example the combination *LSTAT* = $[9.1, 14.66]$ and *INDUS* = $[0.46, 1.47]$. The top-left corner plot presents the error estimated distribution of all 4 models given the total data set. This bivariate MEDP indicates that all 4 models have a considerably worse estimated performance for extremely low values of *LSTAT* ($[1.73, 4.7]$) together with extremely high values of *INDUS* ($[13.89, 27.74]$), showing an interaction between these two variables that cannot be observed by the analysis of each variable MEDP separately (pictured in Figure 3.17). Another interaction where a worse performance is observed concerns central values of *LSTAT* ($[9.1, 14.66]$) with low values of *INDUS* ($[1.52, 3.24]$), but only for the Support Vector Machine, the Random Forest and the Gradient Boosting Machine.

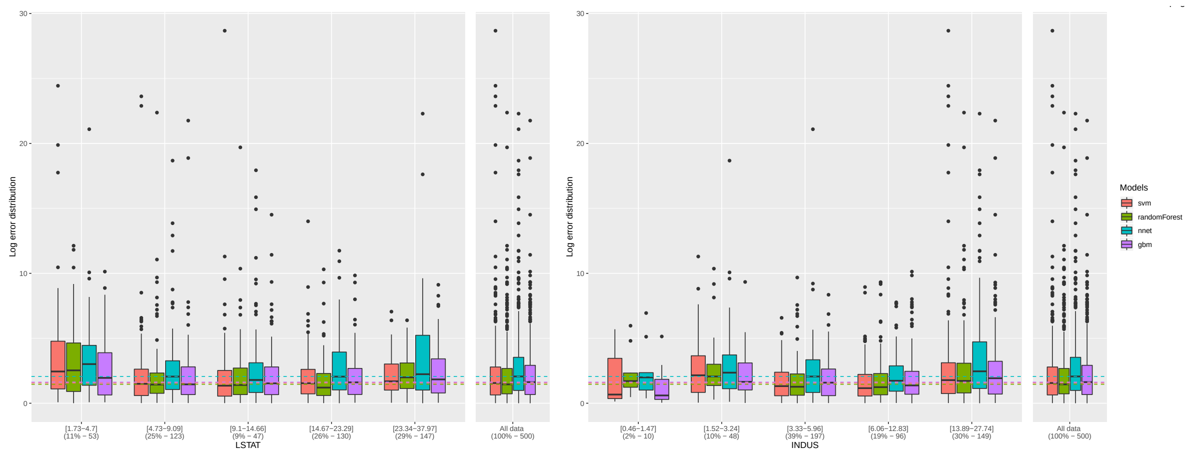


Figure 3.17: MEDP of features *LSTAT* and *INDUS* of data set *boston*.

Figure 3.18 shows another example of a Bivariate MEDP, for features *pgain* and *motor* of the data set *servo*. The top-left corner depicts the MEDP for the categorical feature *motor*, with the large range of error distribution for the Support Vector Machine and the Random Forest when *motor* $= A$ and *motor* $= B$ indicating the existence of interactions with other predictors. In fact, when conditioning for values within *pgain*, the spread in the error distribution is explained: for *pgain* $= 3$ and *motor* $= A$ or *motor* $= B$, the estimated logarithmic error presents significantly higher values for the SVM as well as the RF model, but when *pgain* is within $[5, 6]$, the estimated errors are significantly lower. This example thus illustrates the power of the Bivariate MEDP in uncovering interactions between variables when analysing the performance of a black-box regression model.

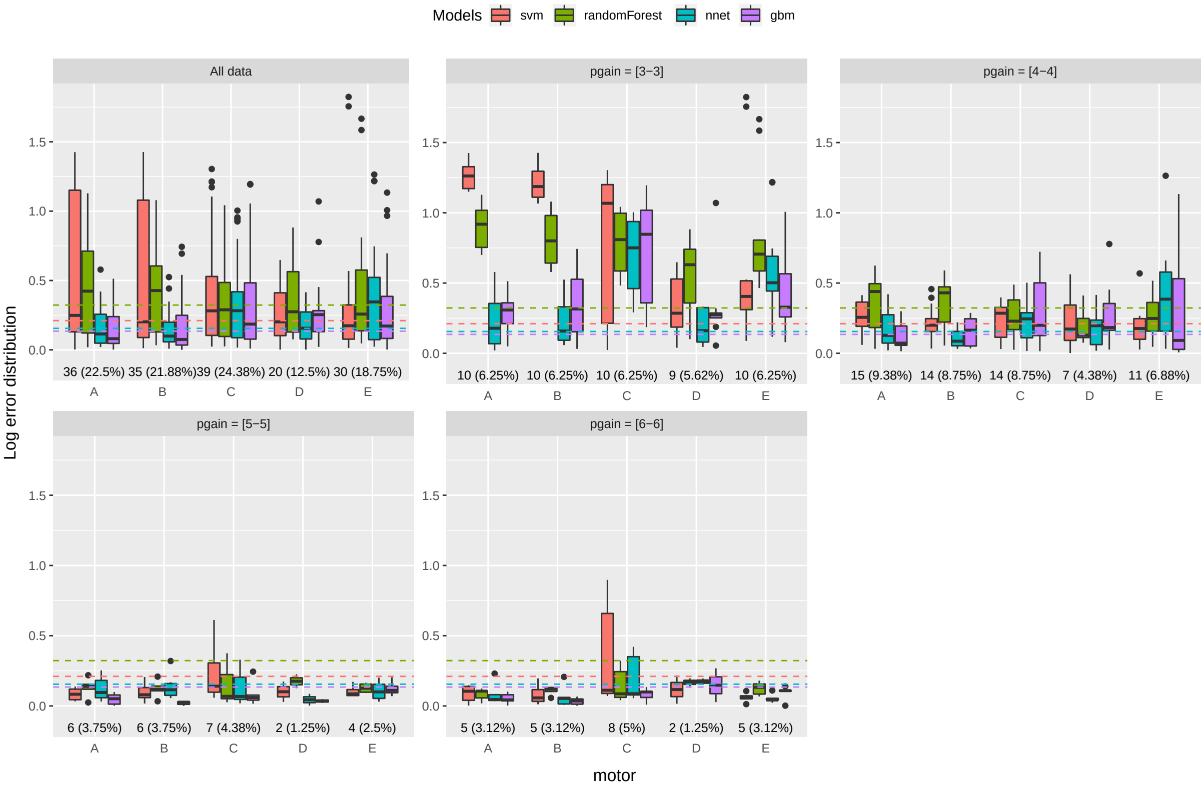


Figure 3.18: Bivariate Multiple model Error Dependence Plot for features *motor* and *pgain* of data set *servo* with logarithmic scale.

Chapter 4

Case Study: Large Scale Marine Protected Areas and Global Fishing Fleets

Our motivating problem, formally defined in Section 4.1, tackles the investigation on how certain Large-Scale Marine Protected Areas (LSMPA) key characteristics, as well as geographic-influenced factors impact fishing effort within and near a LSMPA.

To this purpose, we perform a comprehensive study on predictive models trained with the data set introduced in section 4.2, using methods previously introduced in Section 2.4 and proposed in Chapter 3. In Section 4.3, three distinct models are presented and compared in terms of performance. Subsequently, a single model for each Marine Protected Area is chosen for further evaluation in Section 4.4. Ensuing interpretations of the relation between the predictors and the fishing effort are contained in Section 4.5.

4.1 Motivation

Marine Protected Areas (MPA) have been designated around the world's oceans to enhance global marine protection as to counteract threats originated by overfishing, coastal development and climate change [79, 80]. Large-Scale Marine Protected Areas (LSMPA) encompass all MPAs with an area over 100 000km^2 [81] and should be *actively managed for protection across the entire geographic extent of the area* [82], since these are often located in remote areas, covering a wide range of habitats and various ecosystems.

Established and promised LSMPAs will soon constitute 95% of the global marine protected area, in a total of 25 million km^2 [83], but their fairly recent nature calls for a reevaluation of some ecological and socio-economic factors as initially perceived for smaller MPAs. In fact, despite the restriction or prohibition of fishing activities in these areas, there is a scarcity of information on how LSMPAs interact with surrounding fisheries. Thus, the object of this case study are these large-scale MPAs, due to their novelty as well as their impact on far-ranging fish species.

With this aim, we analysed how a variety of factors influence the fishing effort in thirteen MPAs with an area over 100,000km^2 established before 2015, with key characteristics described in Table B.1 of Appendix B.

This study investigates the area inside and around each MPA, evaluating regions within a radius of up to 500km from the border. This distance reflects a trade-off between the ranges of different fish species (swimming and interaction distances) and the need to keep the data manageable. Analysing areas outside the MPA allows for an insight into the influence of the MPA in neighboring zones, supporting the investigation of possible spillover phenomenons (*'net emigration of fish from protected area'* due to the larger abundance in the MPA generated by the decreased fishing pressure [84]).

The first phase of the study aims to select the most suitable algorithms and parameters that model the system. Hence, we will use methods to analyse the performance of complex regression models, namely our contributions proposed in Chapter 3.

Nonetheless, accountability methods do not provide the entire information required to fully understand the problem. Therefore, in a second phase, some selected state-of-the-art interpretability methods will be applied to discover which factors influence fishing within and near LSMPAs, using the global analysis tools introduced in Section 2.4. Local methods were not adopted since the main purpose is to obtain an overlook of the influence of a variety of factors behind fishing effort patterns, without any particular interest in explaining specific predictions.

4.2 Data Set

The data set for this case study was provided by Dr. Kristina Boerder, a postdoctoral fellow at Dalhousie University (Halifax, Canada). The target variable we are interested in is the fishing effort, which was obtained using Automatic Identification System (AIS) vessel tracking data from 2015 to 2017 from Global Fishing Watch [83]. This information uses the sum of the fishing hours from 2015 to 2017 in particular areas (with size 0.1 by 0.1 degree in latitude and longitude) as the target variable, identifying each instance by the corresponding latitude and longitude of the centroid. 2015 is established as the cut-off year since the AIS tracking data is only usable from this point on. Moreover, the areas in study broaden 500km out from each LSMPA boundary, besides the realms of each LSMPA.

The predictor variables range from environmental parameters to physical and economic characteristics related to each individual LSMPA, and are all numerical:

- Environmental
  - Sea Surface Temperature (SST), *SST*: measured in Celsius;
  - Ocean Productivity, *ave_C*: conveyed by carbon production rates, measured in mg/C/m^2/day;
- Physical

  - Depth, *depth*: in metres;
  - Distance to MPA boundary. *dist_MPA*: in metres;
  - Distance to High Seas, *dist_HS*: in metres;
  - Size, *size*: area of the LSMPA, in km^2;
  - Shape, *shape*: ratio between area and boundary length;
  - Percentage of buffer zone in high seas, *perc_buffer_HS*: percentage of areas near MPA located in high seas with enhanced conservation policies;
  - Percentage of MPA bordering the high seas, *perc_MPA_border_HS*;

- Economic
  - Age, *yrs_since_design*: age of the LSMPA before 2017, measured in years;
  - GDP of designating country, *GDP*: average GDP from 2015 to 2016, in million US$;
  - Enforcement level, *enforcement*: scaled from 1 to 3;
  - Number of management zones, *number_zones*;
  - IUCN number of protected categories; *num_prot_stat_cat*;
  - Percentage of no-take, *perc_NT*: percentage of area with strictly no-fishing policies.

There are only two time-varying predictors: *SST* and *Ocean Productivity*. Both features had then the information aggregated into mean values (*SST* annually from 2009 to 2013 and *Ocean Productivity* monthly from January 2015 to May 2017). The 5 first predictors (*SST*, *Ocean Productivity*, *Depth*, *Distance to MPA boundary* and *Distance to High Seas*) are variable within each MPA, presenting a certain value for a given geographic localization. However, the remaining parameters are fixed and relative to each particular LSMPA, and are defined in Table B.1 in Appendix B.

As we have an interest in understanding both the relation between the fishing effort with all the predictors and the variability for each particular LSMPA, we will analyse these influences for the entire data set and for each individual LSMPA. In sum, our work will focus on 14 data sets: one corresponding to the overall data set provided and another 13 unique for each LSMPA in study. For the analysis of each LSMPA individually we will examine the importance of environmental and physical factors (*SST*, *Ocean Productivity*, *Depth*, *Distance to MPA boundary* and *Distance to High Seas*) whereas the overall analysis will relate all the predictors mentioned previously.

An exhaustive analysis of each of the 13 LSMPAs is extremely extensive to be fully represented here. In light of previous investigation, Dr. Kristina Boerder expressed a special interest in inspecting the feature *%MPA bordering High Seas*, as it is considered an important factor to account for when defining a new MPA, since it strongly influences fishing activities. This was even previously observed by plotting the movement of fishing vessels, which are usually more concentrated along the boundaries between countries waters (country Exclusive Economic Zone) and the high seas. A main justification for this can be provided by the fact that there is no prohibition in fishing within the high seas, while Exclusive Economic Zones (EEZs) only allow fleets from certain nations to engage in fishing activities.

For this reason, in this chapter our focus goes towards two specific LSMPAs, apart from the global analysis, as we will mainly address the Marine Protected Areas that present extreme values of the percentage of MPA bordering High Seas: Galapagos Marine Reserve (MR) and Chagos MPA, with 0% and 81,9%, respectively. The full plots concearning the remaining MPAs can be consulted in the web page `https://github.com/inesareosa/MScThesis`.

Data Set	*Inst*	*Pred*	*Data Set*	*Inst*	*Pred*
Argo-Rowley Terrace	2249	5	**Chagos**	12737	5
Coral Sea	16800	5	**Galapagos**	3333	5
Great Barrier Reef	7226	5	Lord Howe	10288	5
Macquarie Island	807	5	Marianas Trench	14030	5
NP of the Coral Sea	27279	5	Norfolk	8384	5
Pacific Remote Islands	39307	5	Papahanaumokuakea	11297	5
Phoenix Islands	13209	5	**All**	166946	15

Table 4.1: Data sets used for the case study ($Inst$: number of instances; $Pred$: number of predictors). In bold are identified the data sets in focus.

Note that larger LSMPAs generate more data, since the target values and some predictors were obtained geographically. This might influence the global analysis, in which the information from all the LSMPAs is put together, since a larger representation might lead to a stronger impact in the results, as these were not balanced.

4.3 Choosing a Predictive Model

Machine Learning models are widely used to mathematically represent patterns behind a real-world phenomenon. Each machine learning model is based on distinct mathematical foundations and thus it is expected that one might outperform the others in terms of accuracy.

With the aim of investigating how each characteristic and parameter of an MPA relates to the fishing effort, we initially compared the accuracy of 3 algorithms: Support Vector Machine (SVM) [15], Multivariate Adaptive Regression Spline (MARS) [18] and Random Forest (RF) [14].

In order to achieve satisfactory performance for all algorithms, the parameters of each model were tweaked between the values described in Table 4.2 for all the 14 data sets. The experiments were carried out with the R package *performanceEstimation* [85], using Cross Validation to obtain estimates of the performance for each variant.

Algorithm	Parameters	R package

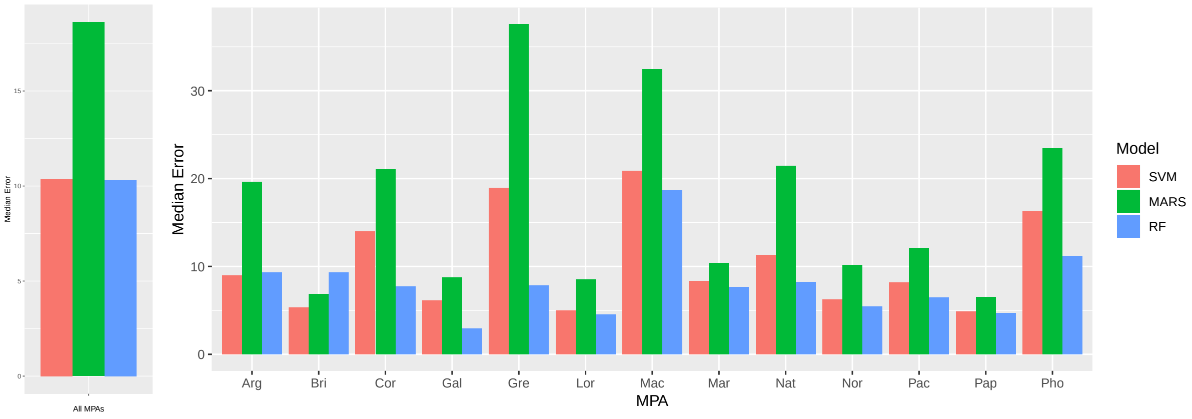
MARS	degree = {1, 2, 3}
	nprune = {2, 26, 51, 75, 100}
		earth [86]

SVM	kernel = {linear, polynomial, radial, sigmoid} cost = {0.01, 0.1, 1, 10}	e1071 [29]

RF	ntree = {250,500,750,1500,300}
	mtry = {2, 3, 4}
		randomForest [30]

Table 4.2: Algorithms used for comparing models, with respective parameter variants and R packages.

After the selection of the model parameters producing better results, the error estimates of each data instance were calculated using 10-fold CV, as described previously in Section 3.2, to enable the evaluation of the model performance.



(a) for All MPAs

(b) for each MPA

Figure 4.1: Median Estimated Absolute Error of each MPA for all three models. *Arg*: Argo-Rowley Terrace MP, *Bri*: Chagos MPA, *Cor*: Coral Sea MP, *Gal*: Galapagos MR, *Gre*: Great Barrier Reef MP, *Lor*: Lord Howe MP, *Mac*: Macquarie Island MP, *Mar*: Marianas Trench MNM, *Nat*: Natural park of the Coral Sea, *Nor*: Norfolk MP, *Pac*: Pacific Remote Islands MNM, *Pap*: Papahanaumokuakea MNM, *Pho*: Phoenix Islands.

Figure 4.1 compares the expected median absolute error for the three algorithms in respect to the global analysis and to each individual MPA. For all MPAs, we can observe that the difference in the median expected error between the SVM and the RF is almost indistinguishable to the naked eye. In fact, the values found (10.365 for the SVM, 10.309 for the RF and 18.614 for the MARS) show the small discrepancy between the first two, when comparing with the MARS model.

Considering each particular MPA, Figure 4.1 shows that the RF has a lower median expected error in 11 of the MPAs, while the SVM presents the best expected median for 2 MPAs (Argo-Rowley Terrace and Chagos MPA). Nevertheless, as for the all MPAs case, in some MPAs the difference between the median expected error is not quite meaningful.

As defended previously in Chapter 3, using a single metric for quantifying an overall performance might conceal some particularities in respect to certain values of the domain. Furthermore, the similar median expected error between the Random Forest and the Support Vector Machine demand further investigation for sensibly selecting the most adequate model. Hence, we employed MEDPs across all predictor values. The selection of the boundaries for each predictor bin could not be made empirically since these are relative to each particular LSMPA, thus requiring very specific expertise. For this reason, we have used the quantiles of the values, segmenting each predictor into 5 bins, comprising extremely low values, low values, central values, high values and extremely high values.

Figure B.1 in Appendix B shows the MEDPs for all predictors when analysing all LSMPAs. We can conclude that for most features the expected performance tends to behave similarly across the values of the domain, as it is example the MEDPs of *SST*, *Ocean Productivity*, *Distance to High Seas*, *Distance to MPA*, *Age*, *Size*, *Number of Protection Categories*, *Enforcement*, *Percentage of Buffer in High Seas* and *Number of zones*. We can infer that the model performance is not related to the values of these predictors.

The MEDPs also show that the estimated performance improves with the increase of *Depth* and *Shape*, being that for the former predictor we can actually establish some differences between the SVM and the RF, since the SVM outperforms for cases when depth is of 0m, while the RF outperforms for values of depth superior to 6000m. The MEDP for feature *GDP* also shows some variability, with the SVM and the RF showing an improved estimated error distribution for countries with high or extremely high GDP. Furthermore, cases with extremely high percentages of no-take percentage are expected to underperform, in spite of the chosen algorithm.

MEDPs disclose consistent underperformance of the MARS model, independently of the operating domain, although not being conclusive when comparing the RF and the SVM. We then proceed from the global to the particular, analysing the MEDPs of each single LSMPA, where we could attain different results. This variability in results might be an indicator that the variability of the performance within the total MPAs is a product of a joint analysis of areas with distinct basal conditions.

Here, as stated before, we will mainly focus on two particular LSMPAs: Chagos MPA and Galapagos MR.

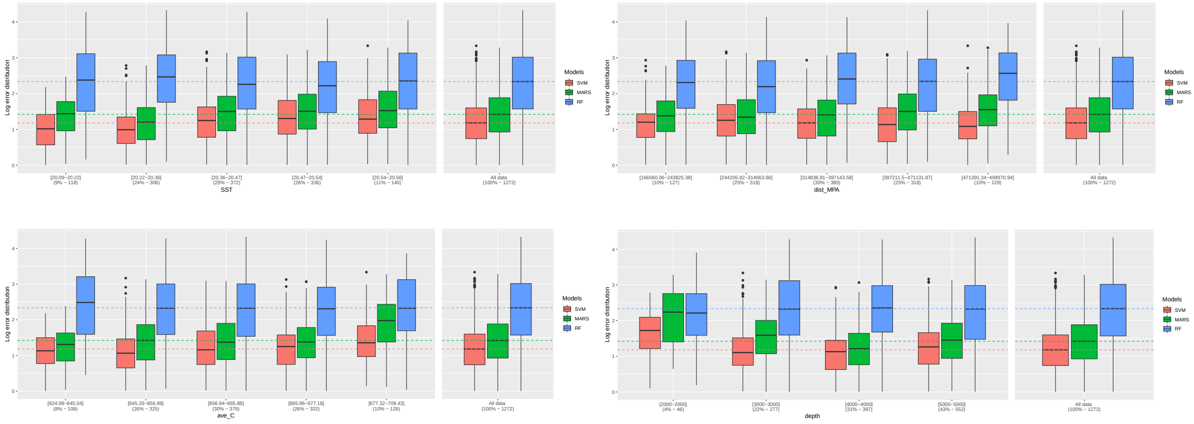
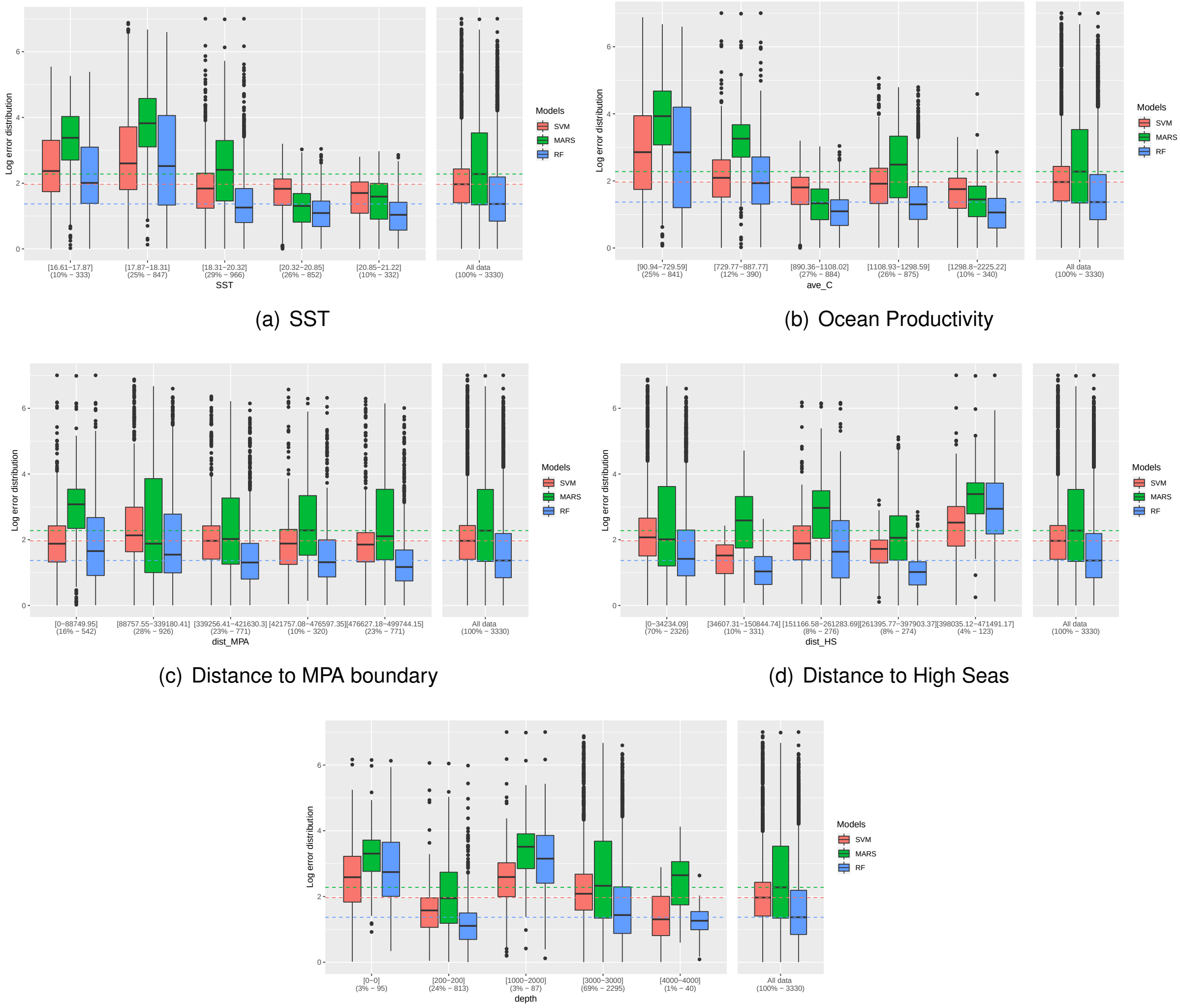


Figure 4.2: Multiple model Error Dependence Plot for Chagos MPA, in respect to each predictor variable values. *Distance to High Seas* is not plotted because all data instances have this parameters with value 0m, and that results in a single bin, equivalent to the one presented for comparison.

The overall estimated error distribution for Chagos MPA, displayed in the rightmost part of each plot in Figure 4.2, indicates a better performance for the SVM when comparing to the other algorithms, with the RF presenting a far worse estimated performance. This inference remains true throughout the bins of all predictor variables values, as it can be concluded by observing Figure 4.2. Furthermore, performance behaviour does not vary across the values of the predictor variables, apart from the cases in which *Depth*=2000m, where both the SVM and the MARS model will underperform when comparing to other depth values.

The fishing effort model for the Galapagos MPA, on the other hand, has the best estimated performance when modelled with a RF, with the MARS model showing a lot of variability in the overall



(a) SST

(b) Ocean Productivity

(c) Distance to MPA boundary

(d) Distance to High Seas

(e) Depth

Figure 4.3: MEDP for each predictor of Galapagos MR.

estimated error distribution, as it can be observed in the rightmost part of all the MEDPs in Figure 4.3.

The MEDP for *SST*, in Figure 4.3(a), shows that all the models improve their performance when applied in higher temperatures. MARS has the worst performance for cases from extremely low to central values of temperature ($[16.61°C, 20.3°C]$), while for higher values ($[20.32°C, 21.22°C]$) is the SVM that underperforms. However, throughout all the range of temperatures, is the RF that has the best performance, as conveyed by the lower median error and the boxplot positioning.

In respect to *Ocean Productivity*, shown in Figure 4.3(b), the tendency is of an improvement in performance with higher carbon production levels. MARS has the higher expected errors for extremely low, low and high values of ocean productivity, while for central and extremely high values is the SVM that underperforms.

The estimated performance seems uncorrelated with the *Distance to the MPA boundary*, since the corresponding MEDP shows a similar behaviour throughout the bins, as can be seen in Figure 4.3(c).

When operating in domains with an extremely high value of *Distance to High Seas* ($[389km, 471km]$), it should be expected an higher error than for all other values of the domain, with a worse performance for the RF and the MARS models (Figure 4.3(d)).

The MEDP corresponding to feature *Depth*, in Figure 4.3(e), shows some cases in which the Random Forest is not the best model in terms of the expected error: when depth is of 0m and between 1000 and 2000m. However, these cases only account for a total of 6% of the data set, which indicates a low prevalence of these depths in and around the MPA.

In light of previous considerations, we can infer that MARS is clearly outperformed by the SVM and RF algorithms for the problem of fishing effort. As the differentiation between the usage of the SVM versus the RF depends on each particular MPA, we opted for utilizing just the Random Forest for coherence when comparing from MPA to MPA. The RF was chosen since it is the most commonly used algorithm in this field as well as it has a better median expected absolute error in most LSMPAs. Thus, from this point forward the investigation will solely utilize the models previously obtained with Random Forests.

4.4 Evaluating the Model

Repeating the overall performance metrics would be redundant as this was already considered beforehand when comparing the models. Moreover, the information shown through the usage of EDPs, which analyses the estimated error distribution across the values of a predictor, was already conveyed when plotting the MEDPs. However, up to this point the interaction between predictors was disregarded.

The usage of PEPs, presented in Section 3.2.4, support an analysis of the performance across the values of all the predictors simultaneously, thus enabling the inspection of possible interactions that might lead to performance degradation.

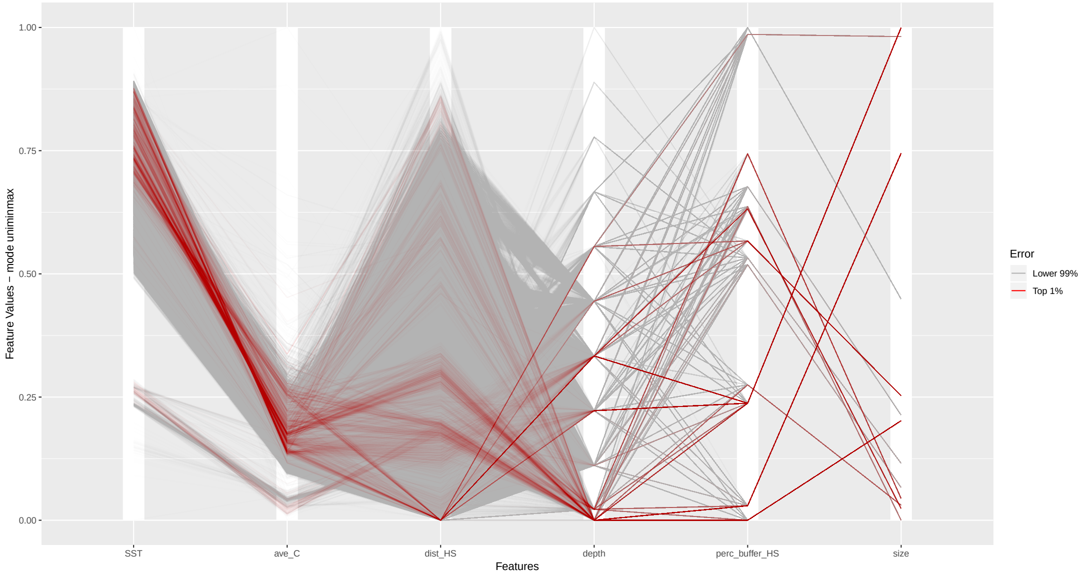


Figure 4.4: PEP for the RF that models all the 13 LSMPAs, for the 6 most important predictors (calculated using PFI).

Figure 4.4 depicts the Parallel Error Plot for the 6 most important predictors (as computed in Section 4.5.1 using Permutation Feature Importance) of the RF that models all LSMPAs. This plot shows that the top 1% of the errors are likely to occur in situations with high temperatures, low values of ocean productivity and extremely low to central depths, not having any particular visible relation with the distance to high seas, percentage of buffer in high seas and size.

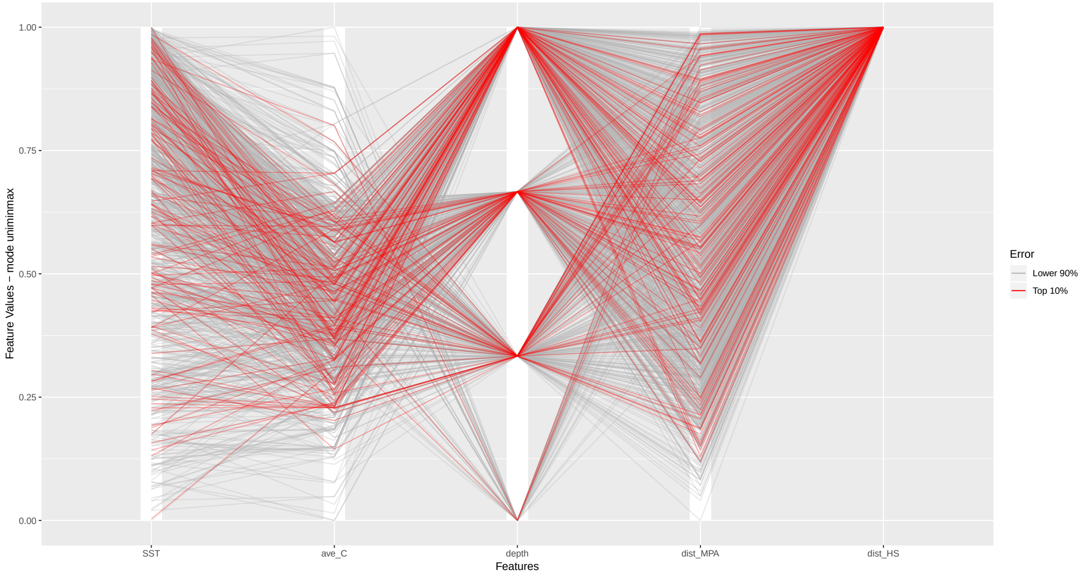


Figure 4.5: PEP of the Random Forest for the Chagos MPA.

The PEP for Chagos MPA, in Figure 4.5, does not show any flagrant correlation between the top errors and the domain of each predictor variable.

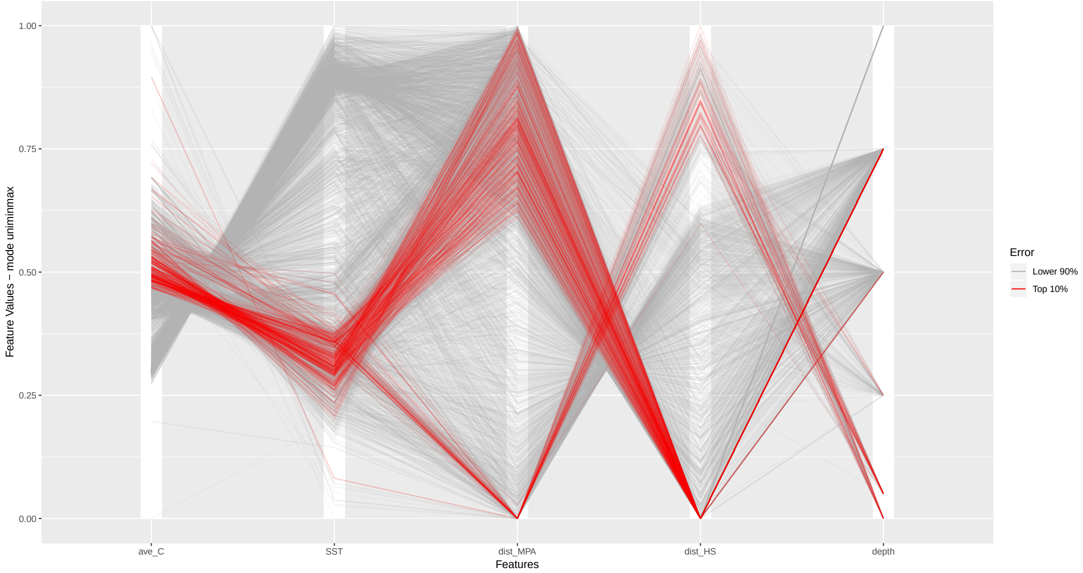


Figure 4.6: PEP of the Random Forest for the Galapagos MR.

In Galapagos MR we could identify some interactions that explain most of the top 10% errors, as observed in the PEP plot represented in Figure 4.6. The Random Forest is expected to underperform for cases with central values of ocean productivity, central/low values of SST, high and extremely low distances to the MPA boundary and extremely low distances to high seas. Inspecting these interactions with the usage of Bivariate EDPs not only were these assumptions confirmed but also it was provided an idea of the range of values with estimated worse performance.

The Bivariate EDPs with interesting behaviour for this particular interaction analysis are represented in Figures B.2 and B.3 of Appendix B. We can conclude that combinations of feature values that lead to an expected worse performance are the following: ocean productivity with high values $[1108.93, 1298.53mg]C/m^2/day$ and SST with extremely low and low values $[16.61°C, 18.31°C]$, extremely low values of distance to high seas $[0, 342km]$ with extremely low values and low values of SST, distance to MPA boundary with central to extremely high values $[339km - 499km]$ and extremely low and low values of temperature as well as with high values of ocean productivity.

We also encountered some interactions that could not be perceived with the usage of the PEP: for cases when depth is of 0 or 200m, in combination with high and extremely high values of distance to high seas $[261km, 398km]$, extremely low distances to the MPA boundary $[0, 89km]$ or high and extremely high values of ocean productivity, the model has a estimated worse performance.

4.5 Global Analysis of the Fishing Effort Models

The aim of this case study is to understand the influence that a set of factors has on fishing effort in areas located inside or nearby LSMPAs. While the adoption of a Random Forest algorithm helps achieving a desirable accuracy and effectively modelling this real-world problem, its inherent opacity obfuscates the comprehension of these relations.

In this section we provide a broad overlook of the models, explaining which factors influence fishing effort the most, as well as how each feature interacts in relation to the target value, resorting to the methods introduced in Section 2.4.

Before interpreting the models *per se*, we believe it is crucial to evaluate the correlation between features, as this information will influence the selection of the algorithms to use. In fact, the correlation plots for each LSMPA, in Figure B.5 of Appendix B attest that almost all MPAs, apart from Macquarie Island, Marianas Trench, Pacific Islands and Papahanaumokuakea, present at least one strong correlation between two features, if using the common guidelines that consider strong correlations as between [0.7,1] and [-1,-0.7] [87].

Most strong correlations are found between *Sea Surface Temperature* and *Ocean Productivity* (5 cases of strong correlations found in the 13 MPAs), as well as between *Sea Surface Temperature* and *Distance do High Seas* (3 cases of strong correlation). In what considers temperature and ocean productivity, we can state these are linked to a certain degree since plankton has optimal temperatures in which it grows better.

The correlation plot for the global analysis, in Figure B.4 (Appendix B), indicate strong correlations between i) *Shape* and *Size*, ii) *Percentage of No-take* and *Age*, iii) *Percentage of Buffer in High Seas* and *Percentage of No-take* and iv) *Number of Zones* and *Number of Protection Categories*

4.5.1 Feature Importance

The Permutation Feature Importance [14], introduced in Section 2.4.1.1, enables the calculation of a score that selects the most important features for a given model, being model-specific for Random Forests. This computation was performed using the R package *randomForest* [30], which returns two distinct values: the % Increase in Mean Squared Error and the Increase in Node Purity, as defined in Section 2.4.1.1.

Features	*% Increase MSE*	*Increase Node Purity*
Sea Surface Temperature	80646.1	397535142.805067
Ocean Productivity	21215.5	294537311.654712
Distance to High Seas	13036.7	257130410.41475
Depth	5825.8	39597424.5
% Buffer in High Seas	4853.1	16805067.6
Size	4792.8	13668884.2
Distance to MPA Boundary	4268.4	191068556.4
Shape	3020.9	13289907.1
Number of Zones	2343.2	13052479.6
% MPA Bordering High Seas	1601.6	5302384.5
Number Management Zones	1408.8	10100464.3
Age	947.1	8010084.4
% No Take	889.8	4553745.8
GDP	679.6	6885640.3
Enforcement	244.2	2775502.5

Table 4.3: Feature Importance for Random Forest modelled from data set with all LSMPAs information.

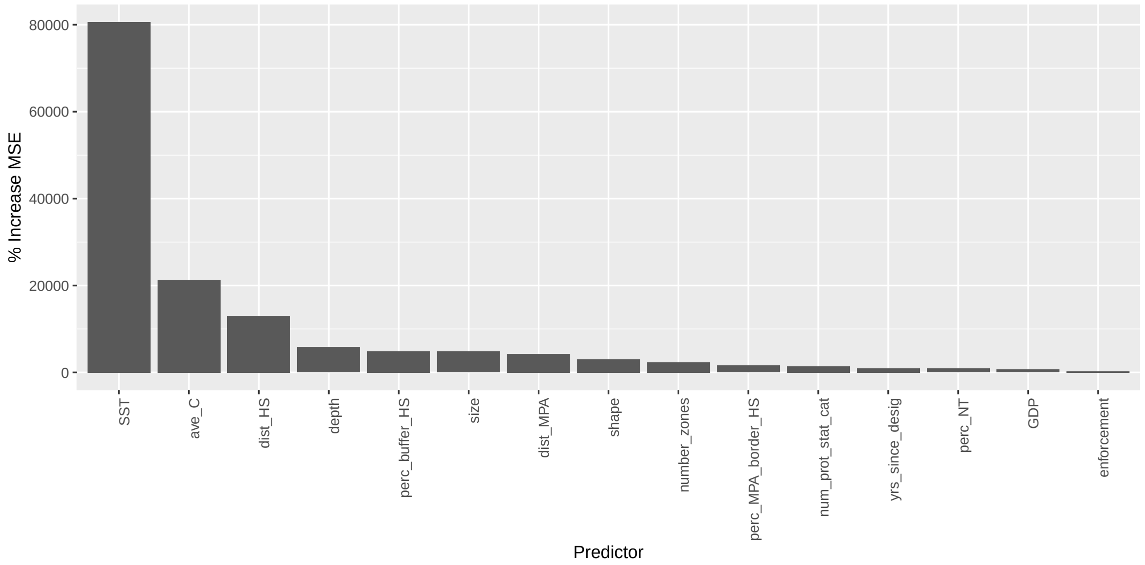


Figure 4.7: Graphic indicating the % Increase in the Mean Squared Error when permuting the values of each feature for predictions using the Random Forest that models all LSMPAs.

Table 4.3 and Figure 4.7 present both measures for the set of all LSMPAs, showing that the primary drivers of overall fishing effort are environmental factors, particularly the SST and the *Ocean Productivity*. This was expected since optimal temperatures lead to high concentration of plankton, which in turn increase the concentration of fish. We emphasize the role of the *Percentage of Buffer in the High Seas*, which is considered with this method as the most important MPA characteristic for determining the fishing effort. This confirms the initial assumption from the domain experts that stated that this factor was one of the most important factors for predicting fishing effort.

MPA	*SST*	*Ocean Productivity*	*Depth*	*Distance to High Seas*	*Distance to MPA*
Argo-Rowley Terrace	**2676.6**	2031.8	1938.5	2091.1	2307.6
Chagos	**196.0**	173.2	102.6	1.72	94.7
Coral Sea	**278886.4**	87124.2	73115.4	72091.1	75559.2
Galapagos	3797.9	**4991.1**	821.7	1015.5	2213.1
Great Barrier Reef	90074.2	94786.3	43805.4	**95848.7the f**	71919.0
Lord Howe	649.0	**1385.4**	324.9	1179.1	410.9
Macquarie Island	25295.5	**32541.6**	22724.9	26393.7	28839.3
Marianas Trench	316.8	202.6	97.3	**327.7**	226.1
Natural Park of the Coral Sea	**57444.3**	8284.4	3957.8	55070.1	12936.0
Norfolk	1408.2	1122.4	250.2	**3081.4**	774.1
Pacific Remote Islands	1188.1	**1329.0**	263.2	702.9	467.0
Papahanaumokuakea	101.5	**194.7**	33.8	155.3	144.8
Phoenix Islands	2892.2	3294.2	426.5	**4749.1**	3119.0

Table 4.4: Feature importance score for the predictors of each individual LSMPA, measured by the unscaled percentage of increase in mean squared error. The most important predictor variable, with higher error increase, is stressed in bold.

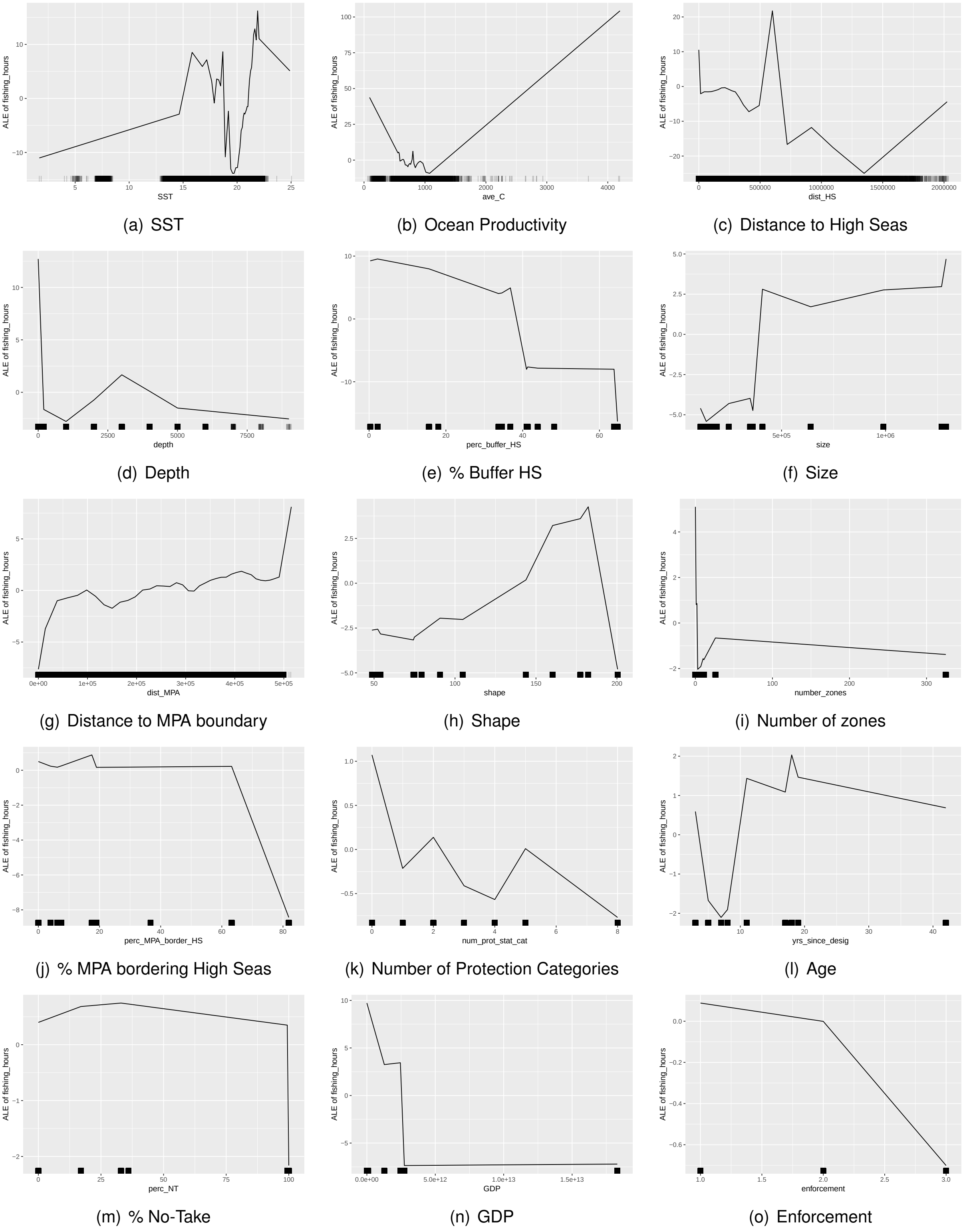
Analysing the relative importance for each of the LSMPAs, with the respective values of the increase in mean squared error represented in Table 4.4, we can conclude that the Ocean Productivity has the greatest influence, top in 5 out of 13 LSMPAs, followed by the Sea Surface Temperature. However, we can observe that in some LSMPAs, as it is example Chagos, the Great Barrier Reef or the Marianas Trench, the difference between the influence of two predictors is quite subtle. Note that the magnitude of these values depends on each individual model and should not be compared between models.

4.5.2 Outcomes with Predictor Variables

To further understand the functioning of the chosen black-box models, it is crucial to determine the type of influence of each predictor variable in the fishing effort patterns. As advocated in Section 2.4.1.2, ALE plots present the most adequate solution to capture this relation when features show some correlation, which is the case for these models. These graphics show how each feature influences the prediction of the Random Forest in average. All plots were obtained using the function *ale* from R package *iml* [51].

Considering Figure 4.8, which pictures the ALE plots of each feature for the global analysis of all thirteen LSMPAs, we can conclude that most variables show clear influence on defining fishing effort patterns. The highest fishing effort values are predicted to be encountered in shallower waters, with temperatures between 15°C and 17°C or above 21°C, as well as for places with very low or very high ocean productivity (below approximately 875$mgC/m^2/day$ or above 1375 $mgC/m^2/day$). Predicted fishing hours tendentially increase with the distance to the Marine Protected Area and show some variability with the distance to high seas, with greater effort in areas located in high seas.

Regarding the parameters specific to each LSMPA, higher predicted fishing effort apparently occurs for older and bigger MPAs with greater shape ratio (largest area per boundary), when designated by countries with a lower GDP, lower enforcement and fewer number of management zones but with stronger protection (conveyed by an higher number of protection categories). Forecasted fishing hours tend to decrease the higher the percentage of buffer located in high seas, and slightly increase with the percentage of MPA bordering high seas, apart for cases with extremely high percentages, where the fishing effort is predicted to be lower. Note that this last assumption might be related solely to the single MPA that has this value (Chagos MPA) and should not be fully trusted for transporting this conclusion to



(a) SST (b) Ocean Productivity (c) Distance to High Seas

(d) Depth (e) % Buffer HS (f) Size

(g) Distance to MPA boundary (h) Shape (i) Number of zones

(j) % MPA bordering High Seas (k) Number of Protection Categories (l) Age

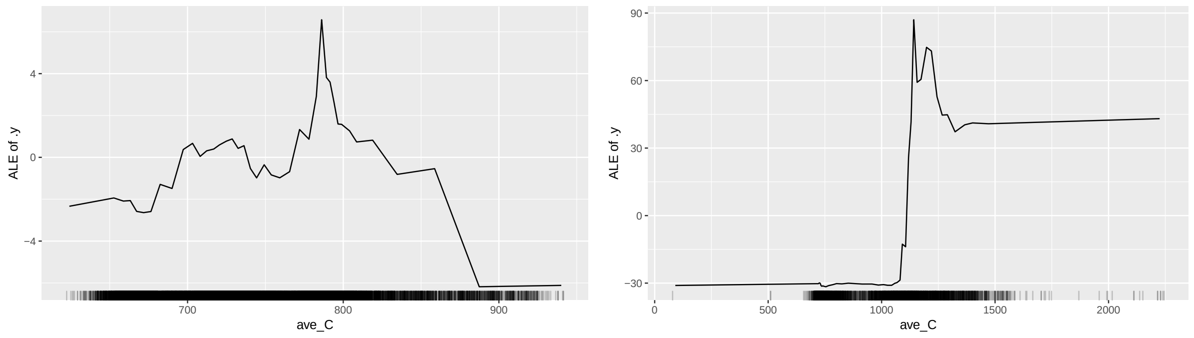
(m) % No-Take (n) GDP (o) Enforcement

Figure 4.8: ALE Plots of the fishing effort in relation to each of the 15 predictors from the Random Forest that models all LSMPAs. The traces in the X-axis indicate the localization of the points in the data set.

other MPAs with similar values.

When analysing the influence of each predictor for each individual MPA, examining the ALE plots that can be consulted in `https://github.com/inesareosa/MScThesis/FishingEffort/ALE`, we can draw that features *Depth* and *Distance to MPA Boundary* show a common directionality through almost all MPAs, coherent with the general assumptions taken for the global LSMPAs. However, some exceptions were noticed: Argo-Rowley Terrace, Coral Sea and Great Barrier Reef show a tendency for lower predicted fishing effort with the increase in distance to the MPA, while the Macquire Island presents lower fishing effort for shallow waters. The other three features (*Sea Surface Temperature, Ocean Productivity* and *Distance to High Seas*) present distinct influence patterns, that seem very unique to the conditions regarding each particular MPA.

Considering the two MPAs in further analysis, Galapagos MR and Chagos MPA, it was set up a comparison feature by feature.

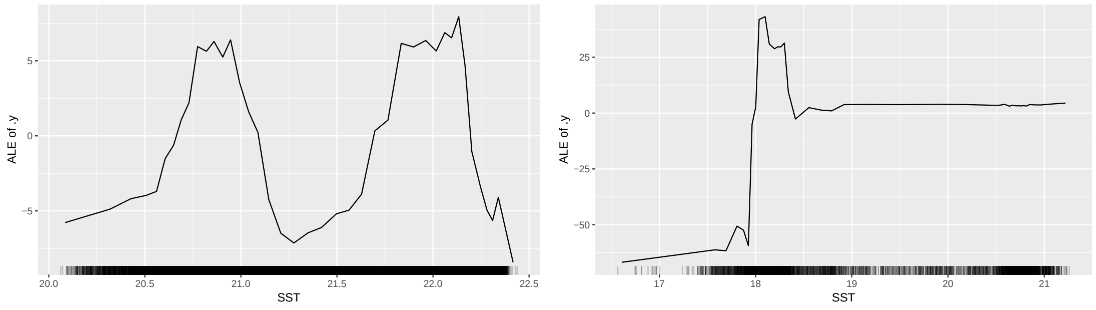


(a) Chagos MPA

(b) Galapagos MR

Figure 4.9: ALE Plot of fishing hours in relation to the distance to ocean productivity, for a single LSMPA. Ocean productivity values expressed in mg/C/m^2/day.

In Figure 4.9, we conclude that central values of ocean productivity lead to higher fishing effort for Chagos MPA, while for Galapagos MR this occurs for higher values of carbon production.

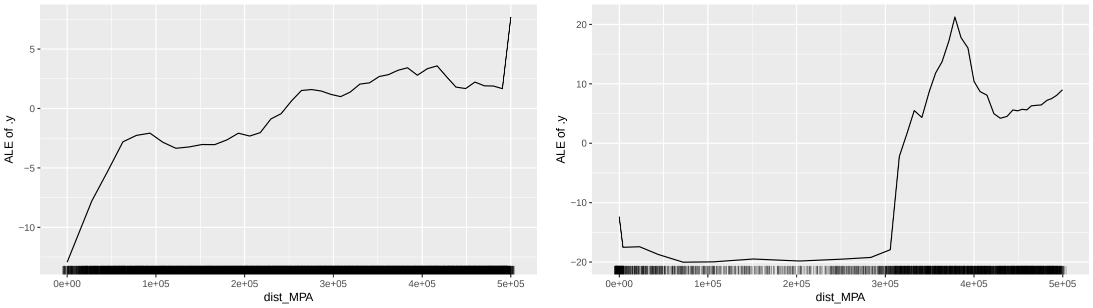


(a) Chagos MPA

(b) Galapagos MR

Figure 4.10: ALE Plot of fishing hours in terms of values of sea surface temperature, for a single LSMPA. SST values expressed in ° Celsius.

Chagos model presents higher fishing effort for temperatures between 20.5°C and 21.2°C or between 21.7°C and 22.2°C, while for Galapagos this occurs for temperatures superior to 18°C. These two behaviours should not be directly compared since the models operate in different ranges of temperature. Nevertheless, according to Dr. Kristina Boerder, this influence should not be overemphasized since the temperature ranges are narrow and thus it is unlikely the presence of a real biological effect.



(a) Chagos MPA

(b) Galapagos MR

Figure 4.11: ALE Plot of fishing hours in relation to the distance to the MPA boundary, for a single LSMPA. Distance to MPA presented in meters.

Fishing hours increase almost monotonically with the distance to the MPA boundary for Chagos, while for Galapagos there is a point from which the fishing effort increases rapidly (around 300km), as seen in Figure 4.11. In the case of Galapagos, this pattern is influenced by the inclusion of an EEZ up to 300km outside the MPA border, as observed in Figure 4.12. The existence of the Ecuadorian EEZ implies that only outside the 300km can international vessels fish, which explains the higher effort.

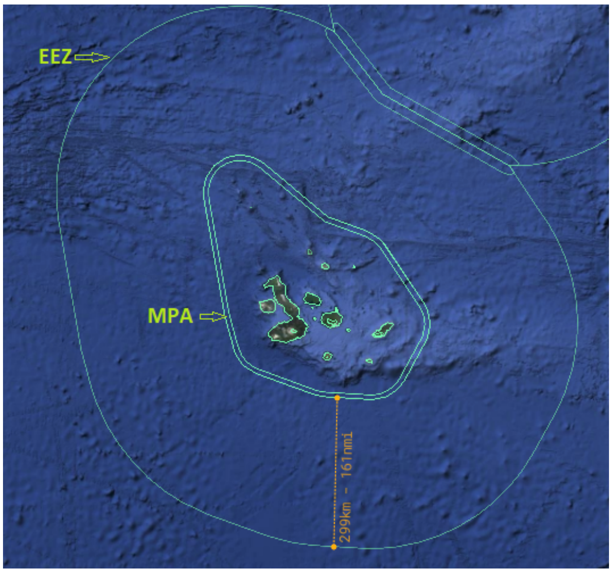
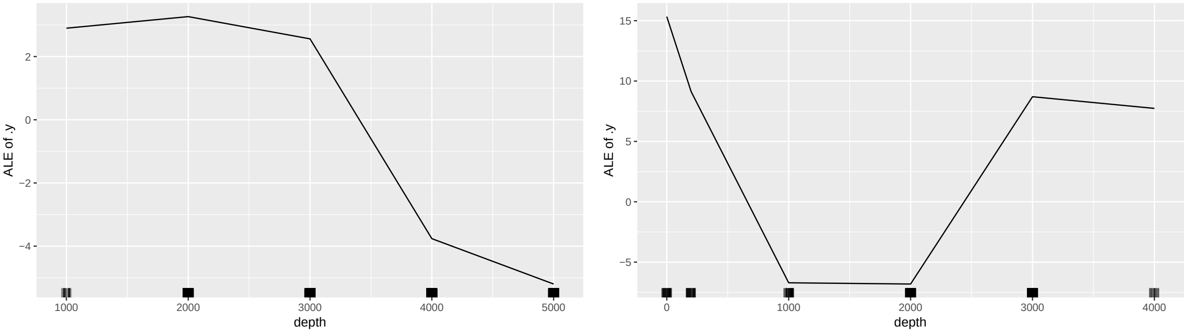


Figure 4.12: Map of the Galapagos MPA and the surrounding EEZ, obtained from the Global Fishing Watch Map [88].

Considering Figure 4.13, we recognize that for Chagos the fishing intensity is predicted to decrease with the depth of the water, while in Galapagos this decrease occurs only up to 2000m, value from which the fishing effort increases again.

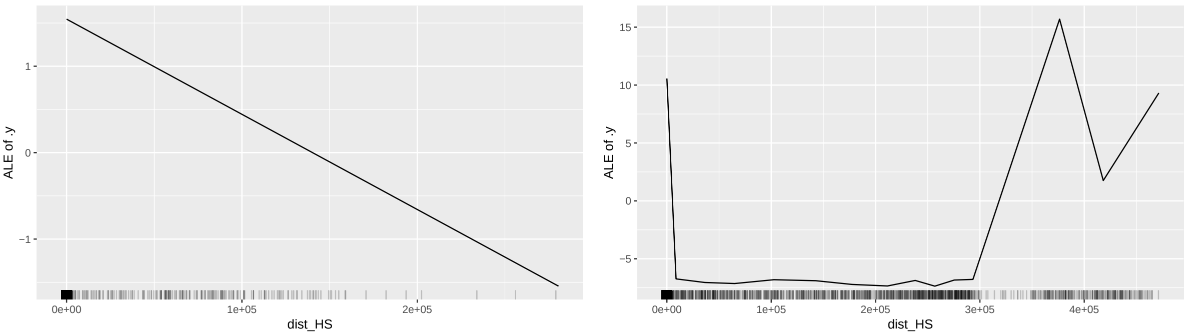
Lastly, as illustrated in Figure 4.14, in Chagos the distance to the high seas establishes a decreasing linear relationship with the fishing effort. The majority of the instances in the data set present a null distance to the high seas, so this might explain the low variability of the fishing effort with the values of this predictor. On the other side, Galapagos present a higher fishing effort in areas located in high seas or areas around 300km away from high seas, which we found in Figure 4.12 to correspond to the



(a) Chagos MPA

(b) Galapagos MR

Figure 4.13: ALE Plot of fishing hours in relation to depth, for a single LSMPA. Depth values expressed in meters.



(a) Chagos MPA

(b) Galapagos MR

Figure 4.14: ALE Plot of fishing hours in terms of distance to High Seas, for a single LSMPA. Distance presented in meters.

boundary of the MPA. The latter increasing effort might be an indication of whether Ecuadorian vessels taking advantage of spillover phenomenons or the presence of a good fishing hotspot.

4.5.3 Interactions

All the algorithms that evaluate the impact of a single predictor in the model outcome ignore interactions between predictors, that can influence the model prediction.

H-Statistic, presented in Section 2.4.1.3, measures the degree of interactions between features, in a scale from 0 to 1, with 0 reporting no interaction and 1 informing that the effect on a prediction arises solely from the interaction. The metrics and corresponding graphics were obtained using R package *iml* [51]. 2D ALE Plots and 2D PDP Plots provide a reasonable solution that evaluates these interactions visually, providing information on which values of the predictor are expected to interact. Here we chose to use the ALE Plots since these are more reliable when operating with correlated features. The respective plots were computed using the R packages *ALEplot* [89] and *iml* [51].

Figure 4.15 plots the H-Statistic for the predictors of the Random Forest that models all 13 LSMPAs. The value for the *Sea Surface Temperature*, close to 1, shows that this feature impacts the outcome almost solely by means of interaction with other features. This is a curious remark if taking into con-

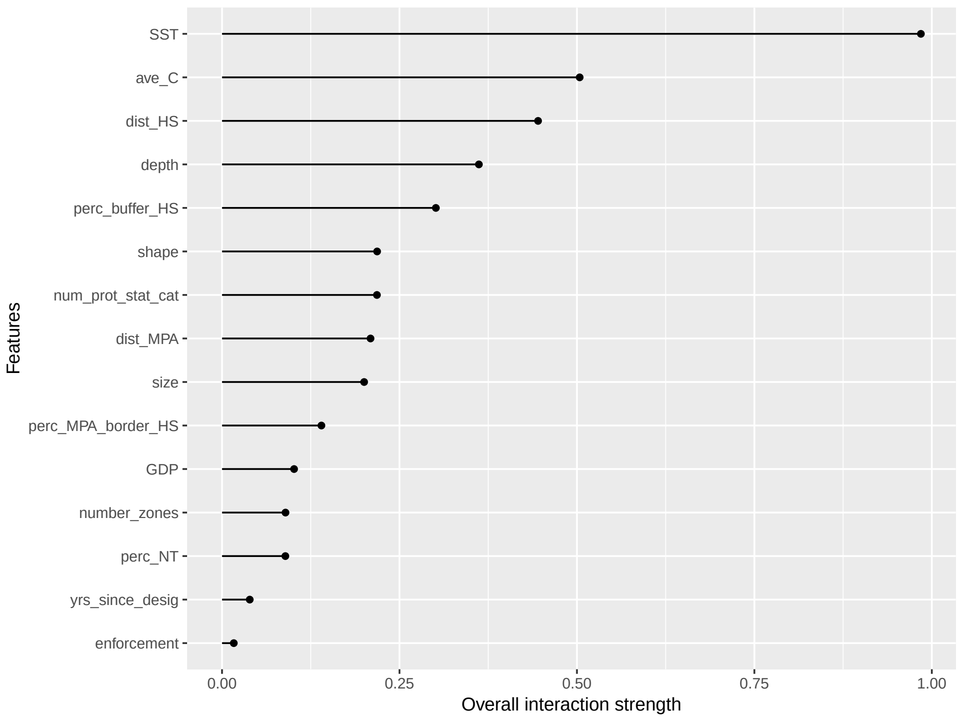
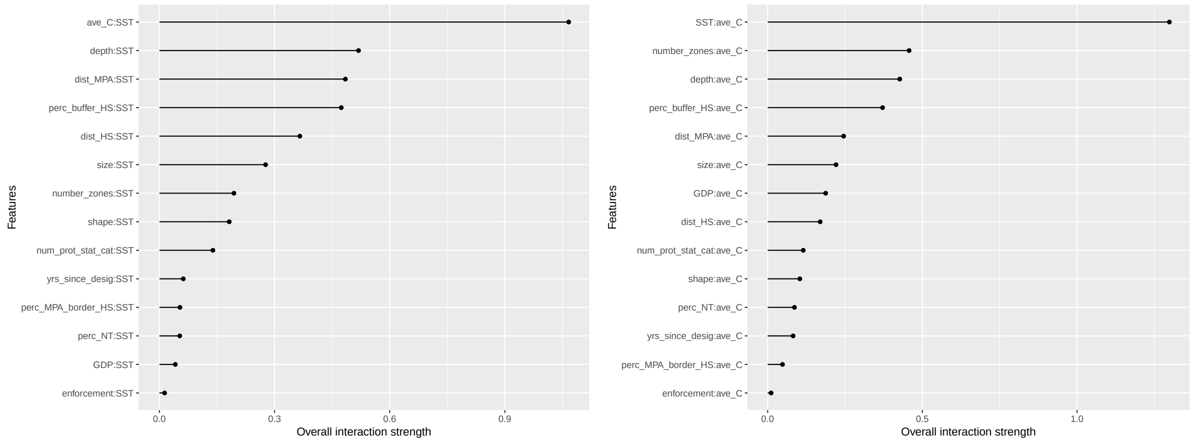


Figure 4.15: H-Statistic for the Random Forest that models all 13 LSMPAs together.

sideration that this was regarded as the most important feature in Section 4.5.1. *Ocean Productivity*, *Distance to High Seas*, *Depth* and *Percentage of Buffer of High Seas* also present a relatively high H-Statistic.



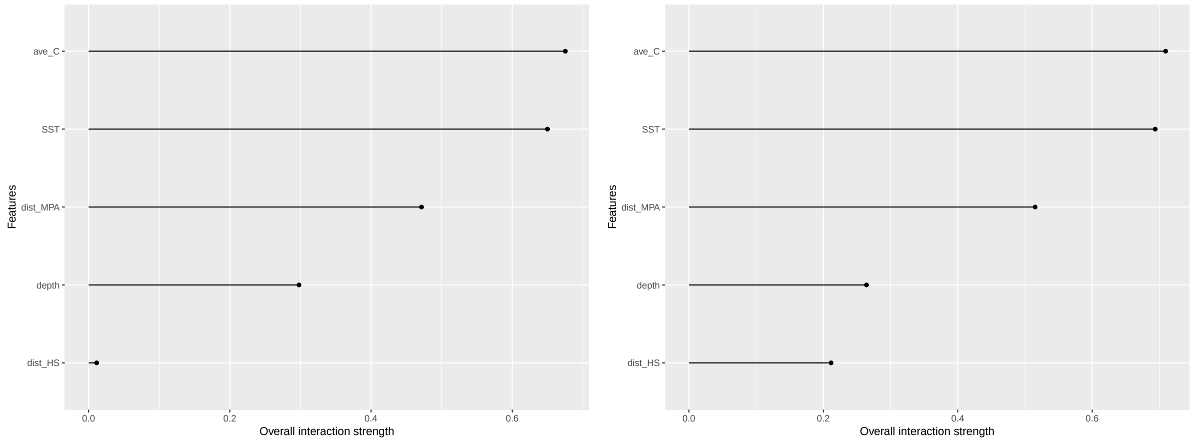
(a) Sea Surface Temperature

(b) Ocean Productivity

Figure 4.16: 2-way H-Statistic for the Random Forest that models all 13 LSMPAs together in respect to a single predictor.

In order to examine in further detail the top 2 H-statistic values, we plot the 2-way interaction H-statistic, to study which feature pairs have the strongest interaction. The highest interaction pair is found for *SST* and *Ocean Productivity*. Note that there is a similarity between the two predictors, since both present strong interactions with two other features: *Depth* and *Percentage of Buffer in High Seas*.

Concerning the two LSMPAs, Chagos MPA and Galapagos MR, with the H-Statistic plots represented

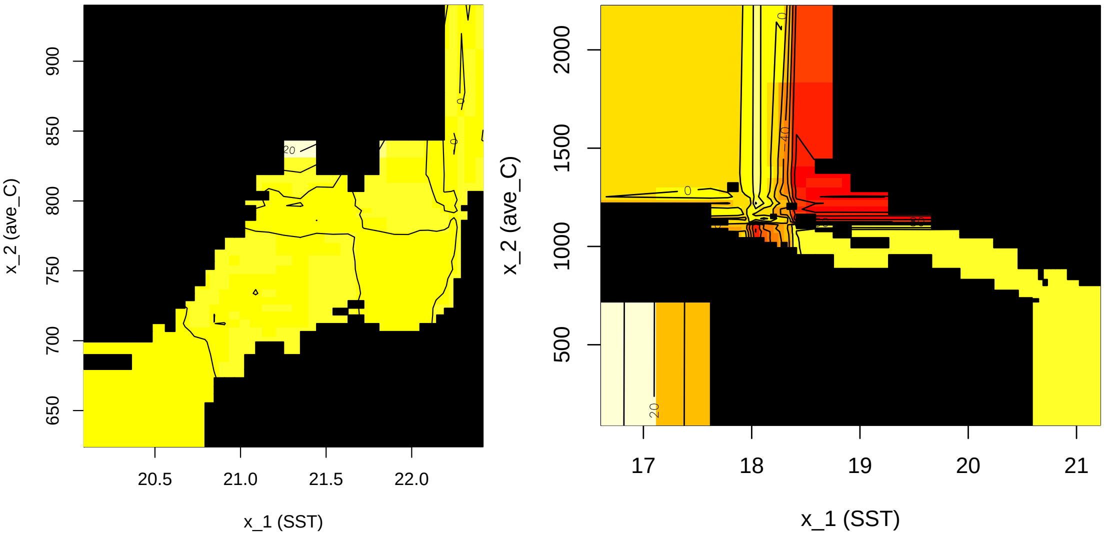


(a) Chagos MPA

(b) Galapagos MR

Figure 4.17: H-Statistic for the Random Forest that models a LSMPA.

in Figure 4.17, both present the stronger interactions for features *Ocean Productivity* and *SST*, following the trend observed in the global analysis. To further investigate these interactions, we employed the 2D ALE Plots for the features with higher H-statistic: *Ocean Productivity* and *SST*, in relation to all other predictors except for *Distance to High Seas* as this feature had a low H-statistic value (almost 0 for Chagos and around 0.2 for Galapagos). Recall that 2D ALE Plots, unlike 2D PDP ALE Plots, only provide information on the values resultant from the interaction of features, and not of the total effect.

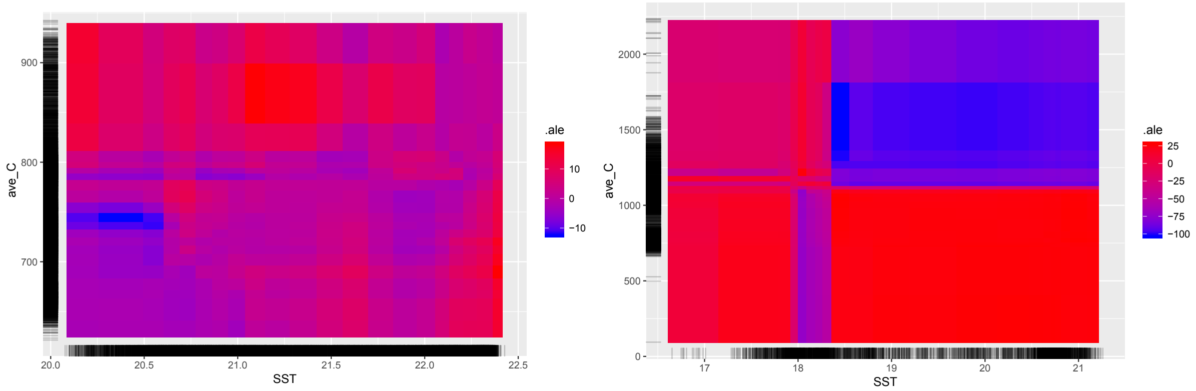


(a) Chagos MPA

(b) Galapagos MR

Figure 4.18: 2D ALE Plot for features *Sea Surface Temperature* and *Ocean Productivity*. Plotted in colour are the areas with points in the data set.

In Figures 4.18 and 4.19 it is plotted the 2D ALE for features *SST* and *Ocean Productivity*. Both depict two approaches for plotting the 2D ALE plot: one showing the interactions for all combinations of values in the domain and another obscuring the part of the graph without any representativeness in the data set. It is quite quaint to observe that the majority of extreme values found in Figure 4.19 actually



(a) Chagos MPA

(b) Galapagos MR

Figure 4.19: 2D ALE Plot for features *Sea Surface Temperature* and *Ocean Productivity*. Plotted all combinations of values, including non-representative areas in the data set. *Red*: High positive fishing effort by interaction; *Blue*: Low negative fishing effort.

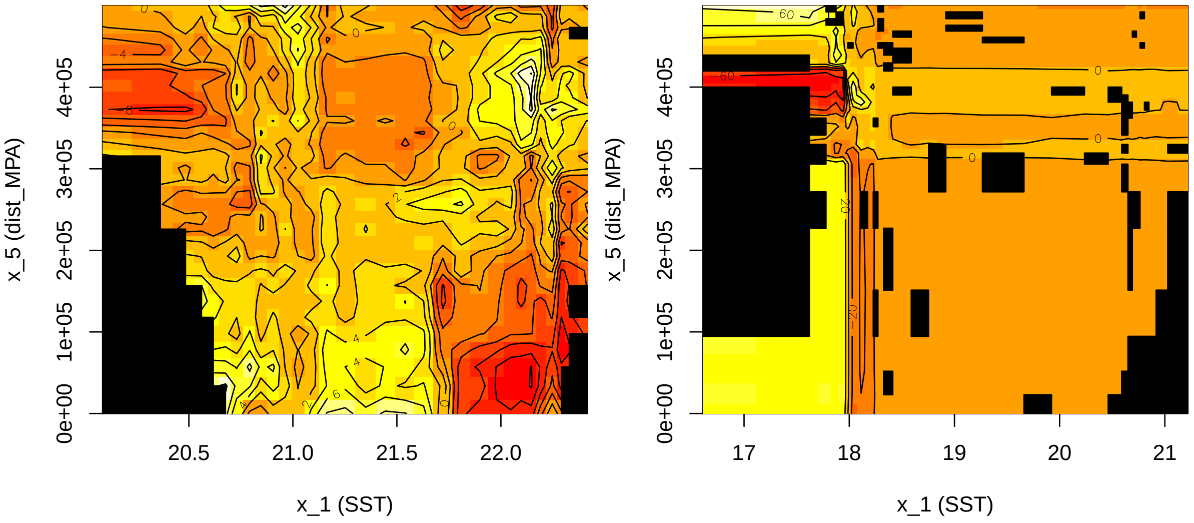
correspond to ”black” areas in Figure 4.18. Therefore, we opted for not analysing the plotting areas not represented in the data set, as this could be misleading and induce into false assumptions, related to improbable combinations of domain values.

For Chagos MPA, we can observe that when the temperature is between 21.3°C and 21.5°C and the carbon production is around 830mg/C/m^2/day there is an interaction that results in an additional prediction of 20 hours of fishing, beyond the expected prediction from each feature individually. Curiously, this is the range of temperatures which lead to a lower fishing effort, if only taking into account the primary effect of *SST* as calculated in Figure 4.10. For the rest of the values it is expected an absence of interactions between both predictors.

For Galapagos MR, the 2D ALE plot is quite more complex, showing plenty of variability resultant from the interactions. A strong positive interaction, of around 20h occurs for temperatures below 17.1°C and carbon production below 750mg/C/m^2/day, which counteracts the primary negative effect of the ocean productivity found in Figure 4.9. On the other hand, negative interactions, reaching -40h and -80h, occur for central temperatures (from 18.4°C to 19.7°C) and carbon production rates superior to 1100 mg/C/m^2/day.

In relation to the interactions between *SST* and *Distance to the MPA Boundary*, plotted in Figure 4.20, Chagos shows additional predictions that range from -8h to 6h. The lowest negative interactions can be expected for low temperatures (below 20.5 °C) and high distances (above 350km) as well as for low distances (<200km) and high temperatures (>21.8°C). High interaction values are found i) near the MPA boundary for temperature between 21°C and 21.5°C, counterbalancing the negative effect of each primary driver within those domains (Figures 4.10 and 4.11), as well as ii) along high distances (around the 400km) and high temperatures (>22°C).

Concerning Galapagos, as seen in Figure 4.20, the interactions occur only for temperatures below 18.3°C, point in which there is a peak of effort driven by the temperature (4.10). For areas closer to the boundary, below 300km, the interaction is estimated to be negative (about -20h) if the is temperature

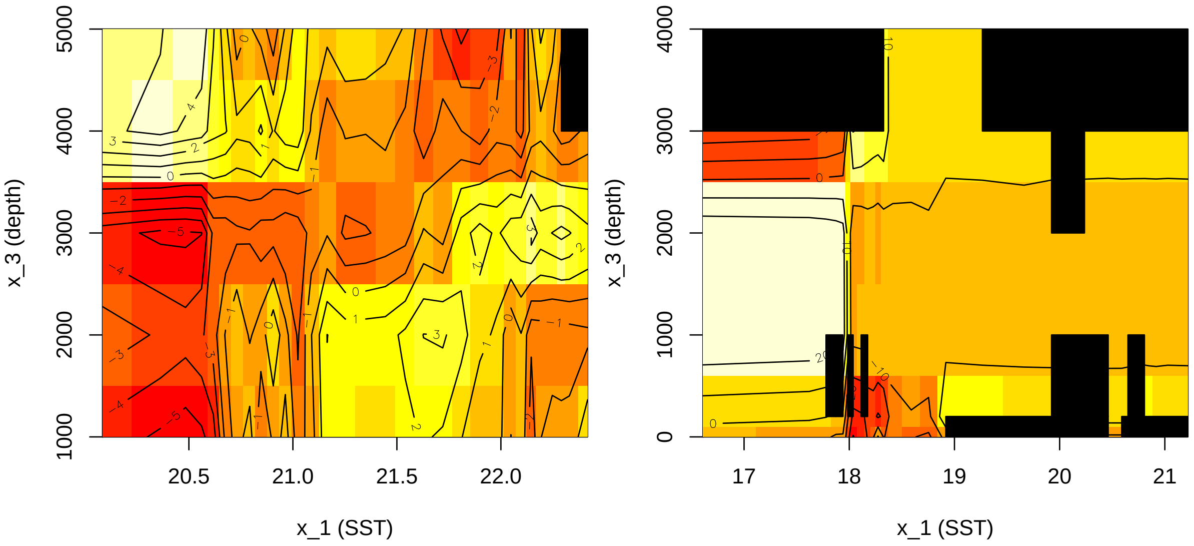


(a) Chagos MPA

(b) Galapagos MR

Figure 4.20: 2D ALE Plot for features *Sea Surface Temperature* and *Distance to the MPA Boundary*.

between 18°C and 18.3°C or positive (about 20h) if the temperature is lower: this shows that the interactions have an effect that counteracts the fishing effort peak and the low effort for low temperatures within the EEZ. Further away from the boundary, for distances of 400km and temperatures bellow 18°C the model has an interaction of -60h, but with the increase of the distance this interaction also increases up to 60h.



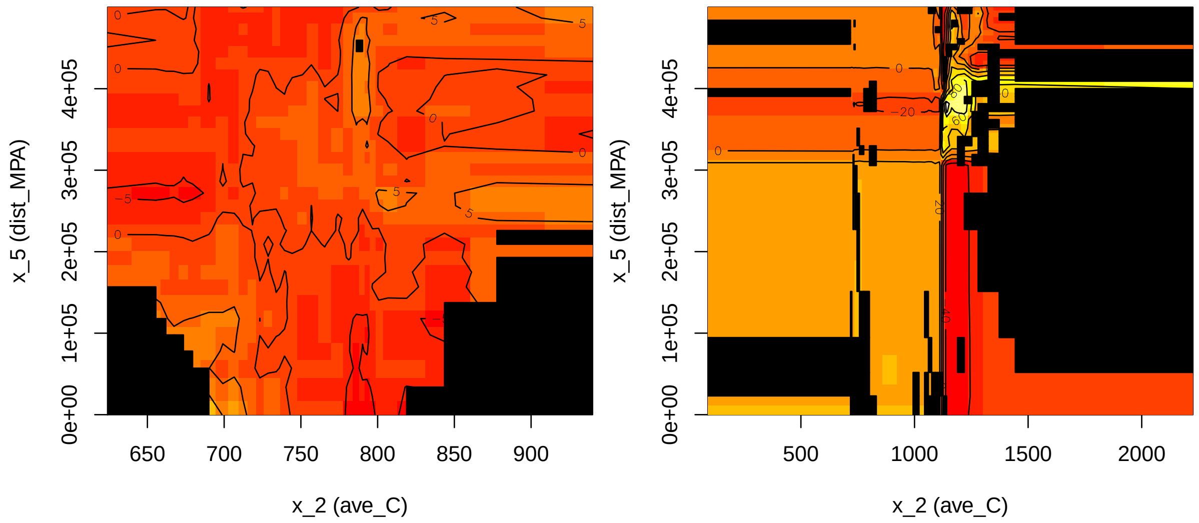
(a) Chagos MPA

(b) Galapagos MR

Figure 4.21: 2D ALE Plot for features *Sea Surface Temperature* and *Depth*.

Figure 4.21 depicts the 2D ALE Plot for *SST* and *Depth*. Chagos plot shows a negative interaction for low temperatures and low depths (SST<21°C and depth<3500m) and positive high interactions for deep areas with low temperatures (SST<21°C and depth>3500m). This indicates that at low temperatures the tendency of the fishing effort decreasing with depth is actually reversed (Figure 4.13).

On the other hand, Galapagos MR shows an estimated high interaction (around 20h) for temperatures below 18°C and depths between 600m and 2500m, which signifies that for low temperatures the negative impact of these depth values (as seen in Figure 4.13) is reversed. Low negative interactions (-10h) are encountered for depths above 2500m if the temperature is below 18°C or for temperatures between 18°C and 19°C for depths under 600m, both cases interestingly with depth values that lead to a positive primary influence in the fishing effort (c.f. Figure 4.13). In this case, as for all other interactions with *SST*, we could observe that the interactions only occur in cases below and around 18.5°C.



(a) Chagos MPA

(b) Galapagos MR

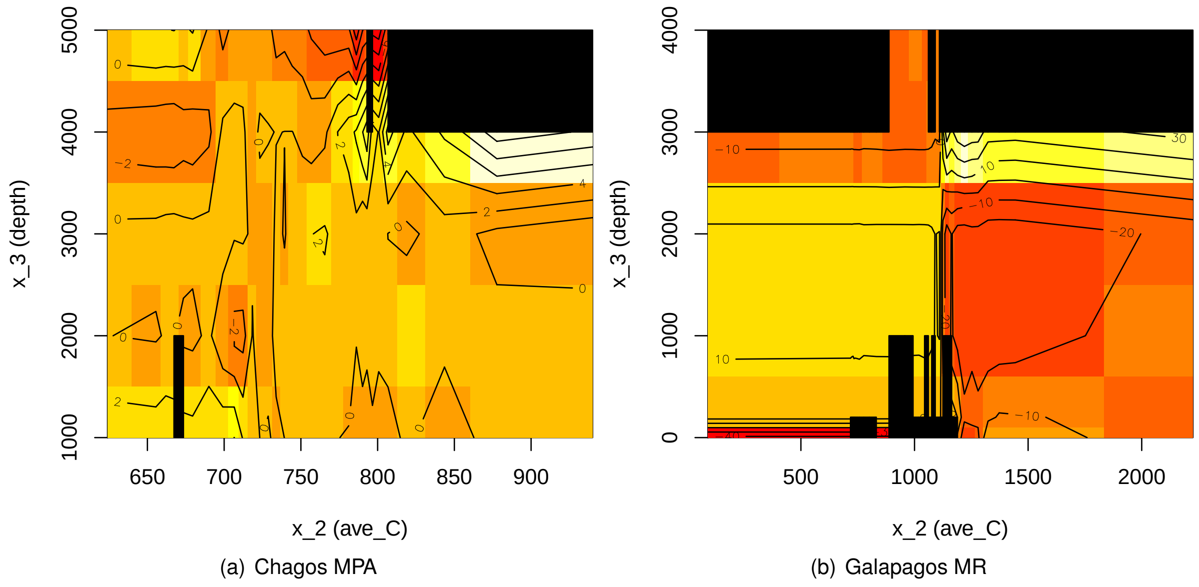
Figure 4.22: 2D ALE Plot for features *Ocean Productivity* and *Distance to MPA Boundary*.

Figure 4.22 depicts the 2D ALE Plot for *Ocean Productivity* and *Distance to MPA Boundary*. Chagos MPA interactions range from -5 to 5 extra fishing hours, but does not denote any particular pattern within the values of the predictors.

Galapagos MR shows that for distances around the 400km, near the fishing effort peak lead by the distance to the MPA as observed in the ALE plot of Figure 4.11, there is a strong positive interaction, around an additional 60h, when the carbon production is around 1200mg/C/m^2/day. However, for lower carbon production this value is negative, of about -20h, strengthening the low fishing effort in regions with low ocean productivity values (c.f. Figure 4.9).

Lastly, Figure 4.23 shows the relation between *Ocean Productivity* and *Depth*. Chagos presents low interactions, from -2h to 2h, apart from two situations: i) for deep areas (>3500m) when the ocean productivity is high for that MPA (>850mg/C/m^2/day), in which there is an interaction of 4h, that counteracts the decrease of effort for high depths (c.f. Figure 4.13), as well as in ii) deep areas (>4500m) with central values of carbon production (around 800mg/C/m^2/day) where is expected a negative interaction, that nullifies the peak of effort observed around those values of ocean productivity, seen in the respective ALE in Figure 4.9.

Galapagos 2D ALE Plot depicts an estimated high interaction (between 10h and 30h) for high carbon



(a) Chagos MPA

(b) Galapagos MR

Figure 4.23: 2D ALE Plot for features *Ocean Productivity* and *Depth*.

production rates in deep areas (ocean productivity>1000mg/C/m^2/day and depth>2500m), intensifying the high effort predicted individually by each predictor with corresponding values of domain (in Figures 4.9 and 4.13). Furthermore, it identifies that zones with low carbon production rates in shallow areas (ocean productivity <1000mg/C/m^2/day and 100<depth<2500m) also have positive interactions. All other cases present a negative interaction.

4.5.4 Summary

The foregoing chapter presents a case study that aims to understand which and how certain factors drive fishing effort in Large Scale Marine Protected Areas. We analyse a set of thirteen LSMPAs, as well as two individual ones: Chagos MPA and Galapagos MR.

Firstly, we began by comparing the performance of three distinct algorithms: MARS, RF and SVM, inspecting both the overall performance as well as the performance with respect to the predictor variables values. The MARS algorithm was shown not to be adequate for this problem, since it underperforms when comparing to RF and SVM. Considering the two single LSMPAs, the Multiple model Error Dependence Plots helped concluding that the performance of the algorithms modelling Chagos does not vary across the range of the predictors, while for Galapagos the performance of the models tend to increase with higher values of *SST* and *Ocean Productivity*. Random Forests present a slightly better median expected error than Support Vector Machines in the overall case and for 11 out of 13 LSMPAs. Thus, henceforth only the RFs were studied.

Parallel Error Plots were then used to investigate the existence of interactions between predictors that lead to performance degradation. For the global RF, the top errors are expected to occur in cases of high temperatures, low values of ocean productivity and extremely low to central depths. Galapagos' RF also showed a clear interaction pattern that affected performance, being expected to underperform for conditions of central values of ocean productivity, central/low values of SST, high and extremely low distances to the MPA boundary and extremely low distances to high seas.

Finally, the Random Forest models were probed in terms of interpretability to inform the end user about the influence of each factor in determining the fishing effort. Both *SST* as *Ocean Productivity* were identified as the factors with most impact in predicting the target value. The role of the *Percentage of Buffer in High Seas* was also emphasized, since it is regarded as the most important MPA characteristic in forecasting fishing effort patterns. This latter result was found to be quite valuable for the domain expert, since it confirmed the initial assumption of the high importance of this factor.

The influence of each predictor was then analysed, by plotting the estimated average change in the fishing hours across the domain values of each predictor using Accumulated Local Effects plots. Features *Depth* and *Distance to MPA Boundary* were found to have a common directionality through almost all MPAs, with fishing effort tending to be higher in shallow waters and to increase with the distance to the boundary. On the other hand, *SST*, *Ocean Productivity* and *Distance to High Seas* have patterns that are very MPA-specific. Regarding Galapagos MR, it is predicted higher fishing effort for situations with higher temperatures or/and ocean productivity, located in the high seas or near the MPA border. Concerning Chagos MPA, the fishing effort is expected to increase with the distance to the MPA border and to decrease with depth, showing an higher tendency for central values of ocean productivity.

In order to obtain the whole picture of the case study, the interactions between variables were also examined. Sea temperature and ocean productivity are known to be linked to a certain degree since plankton has an optimal growth in certain temperatures. As a matter of fact, when calculating the H-statistic, a metric that quantifies the level of interactions between predictors, both features are shown to highly interact when forecasting the fishing effort, specially with each other but also with other predictors. These findings were deemed as quite interesting by the domain expert since these take into consideration the intricate relations between environmental and physical parameters. Lastly, two-dimensional ALE plots for Chagos and Galapagos were depicted to visualize the interactions in relation to the values of two predictors. Moreover, we revealed that for some particular feature combinations, the interactions are shown to nullify the fishing effort peaks observed in the ALE plots of at least one of the predictors.

Chapter 5

Conclusions

5.1 Summary

The increasing complexity of the recent Machine Learning algorithms has intensified the demand for tools that help understanding these models. This work described some methods that help explaining black-box regression models, with a particular focus on *post hoc* model-agnostic visualization tools. We addressed two aspects of explainability: predicting and explaining the performance of a model (accountability) as well as interpreting how each variable influences a model prediction (interpretability).

An overview of the existing methods in both topics was provided, with the introduction of state-of-the-art *post hoc* model agnostic tools, with some being subsequently formulated and compared.

We claim that accountability tools increase the ability to correctly access the risks behind trusting a model. Due to some gaps in the currently existent methods, we described a novel approach, which aims at understanding the factors that may lead to worse performance of the models by relating the expected error to the values of the predictor variables. Under this approach, we proposed the Error Dependence Plot (EDP), which visualises the expected error distribution against the range of one predictor variable (or two/three if it is the Bivariate/Trivariate EDP variant). Then, the reliability of the EDPs was assessed, both visually and formally, by comparing the error distribution predicted from a training set with the actual error distribution obtained with a test set. EDPs were proven to be trustworthy if enough data is available to ensure reliable Cross Validation estimates.

EDPs cannot represent more than 3 predictors simultaneously, which is quite limitative considering that most real world problems have many more features. To address this, we propose an extension of the EDP: the Parallel Error Plot (PEP), which represents the expected error profile for various predictors simultaneously. Thus, this method enables the identification of interactions between features that lead to worse expected performance.

Finally, we proposed a tool, the Multiple model Error Dependence Plot (MEDP), that allows the end user to compare the performance of different black-box models. This method informs on the existence of ranges of values in which a model comparatively outperforms, hence allowing the user to choose the model most suited for the range of predictor values it will operate on.

All proposed tools were formulated and then followed by several illustrative examples applied to distinct black-box models, trained with real world data sets. These use cases depicted the utility of the plots in accounting for the risk associated with the use of a black-box model. Furthermore, all the plots are fully reproducible since the tools were made publicly available.

Lastly, we investigated a case study on the fishing effort on thirteen Large-Scale Marine Protected Areass (LSMPAs). In order to select the most adequate model for the fishing effort in relation to certain environmental factors and key characteristics we used some of the previously proposed accountability tools to compare the performance of three different algorithms. MARS algorithm was concluded to be outperformed by SVM and RF for this specific problem. However, each LSMPA was shown to have a particular performance trend. We then performed a comprehensive interpretative analysis of the chosen predictive models through the usage of some of the *post hoc* visual tools analysed. This study lead to some interesting findings in terms of feature importance, relation of features with outcome and interactions between features. We extensively dissected two particular LSMPA: Chagos MPA and Galapagos MR, since these present the extreme values in percentage of buffer in high seas, which was calculated as the most important key characteristic for predicting the fishing hours in a given region.

5.2 Future Work

It should be taken into consideration that at this present time a clear theory of explainable AI is non-existent. Actually, the definitions on what to consider as explainable are divergent, which consequently deviates the investigation objectives of experts.

Some interesting research directions could be related to the development of a tool that would ally interpretability and accountability characteristics. For instance, combining a feature effect plot with the information of the expected error for a given predictor variable value.

The proposed EDPs are quite useful in determining domain values where the model has a distinct expected error behaviour from the perceived from the overall performance. Nevertheless, investigating these particularities becomes increasingly onerous with the escalation of the number of predictors, specially if the end user is not probing any specific variable. Thus, it would be interesting to automatize the searching process, for instance by comparing the distribution of the errors in each bin of the EDP with the overall error distribution.

In what regards the fishing effort case study, we would like to inspect the models using local methods, such as the SHAP values or counterfactual explanations. These tools could provide an insight on some specific cases, such as determining the reasons behind extremely high predicted fishing effort.

Bibliography

[1] Leilani Gilpin, David Bau, Ben Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. Explaining explanations: An overview of interpretability of machine learning. pages 80–89, 10 2018. doi: 10.1109/DSAA.2018.00018.

[2] Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should I trust you?": Explaining the predictions of any classifier. *CoRR*, abs/1602.04938, 2016. URL http://arxiv.org/abs/1602.04938.

[3] Alex A. Freitas. Comprehensible classification models: A position paper. *SIGKDD Explor. Newsl.*, 15(1):1–10, March 2014. ISSN 1931-0145. doi: 10.1145/2594473.2594475. URL http://doi.acm.org/10.1145/2594473.2594475.

[4] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In Sorelle A. Friedler and Christo Wilson, editors, *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pages 77–91, New York, NY, USA, 23–24 Feb 2018. PMLR. URL http://proceedings.mlr.press/v81/buolamwini18a.html.

[5] Latanya Sweeney. Discrimination in online ad delivery. *Communications of the ACM*, 56, 01 2013. doi: 10.2139/ssrn.2208240.

[6] Sarah Tan, Rich Caruana, Giles Hooker, and Yin Lou. Distill-and-compare. *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society - AIES '18*, 2018. doi: 10.1145/3278721.3278725. URL http://dx.doi.org/10.1145/3278721.3278725.

[7] Jenna Burrell. How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society*, 3(1):2053951715622512, 2016. doi: 10.1177/2053951715622512. URL https://doi.org/10.1177/2053951715622512.

[8] Matt Turek. Explainable artificial intelligence (XAI). URL https://www.darpa.mil/program/explainable-artificial-intelligence. Accessed:1/09/2019.

[9] Inês Areosa and Luís Torgo. Visual interpretation of regression error. *EPIA 2019: Progress in Artificial Intelligence*, 08 2019. doi: 10.1007/978-3-030-30244-3_39.

[10] Inês Areosa and Luís Torgo. Explaining the performance of black box regression model. 2019.

[11] Thomas M. Mitchell. *Machine Learning*. McGraw-Hill, Inc., New York, NY, USA, 1 edition, 1997. ISBN 0070428077, 9780070428072.

[12] Karl Pearson. Note on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London*, 58:240–242, 1895.

[13] C. Spearman. The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):72–101, 1904. ISSN 00029556. URL `http://www.jstor.org/stable/1412159`.

[14] Leo Breiman. Random forests. *Mach. Learn.*, 45(1):18–21, October 2001. ISSN 0885-6125. doi: 10.1023/A:1010933404324. URL `https://doi.org/10.1023/A:1010933404324`.

[15] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, September 1995. doi: 10.1007/bf00994018. URL `https://doi.org/10.1007/bf00994018`.

[16] Brandon Greenwell, Bradley Boehmke, Jay Cunningham, and GBM Developers. *gbm: Generalized Boosted Regression Models*, 2018. URL `https://CRAN.R-project.org/package=gbm`.

[17] W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S*. Springer, 4th edition, 2002.

[18] Jerome H. Friedman. Multivariate adaptive regression splines. *The Annals of Statistics*, 19(1):1–67, March 1991. doi: 10.1214/aos/1176347963. URL `https://doi.org/10.1214/aos/1176347963`.

[19] Sanford Cook, R. Dennis; Weisberg. *Residuals and influence in regression*. New York: Chapman and Hall, 1982.

[20] Norman Richard Draper and Harry Smith. *Applied regression analysis*. Wiley, New York [u.a.], 1966. URL `http://gso.gbv.de/DB=2.1/CMD?ACT=SRCHA&SRT=YOP&IKT=1016&TRM=ppn+022791892&sourceid=fbw_bibsonomy`.

[21] Paolo Tamagnini, Josua Krause, Aritra Dasgupta, and Enrico Bertini. Interpreting black-box classifiers using instance-level visual explanations. In *Proceedings of the 2Nd Workshop on Human-In-the-Loop Data Analytics*, HILDA'17, pages 6:1–6:6, New York, NY, USA, 2017. ACM. ISBN 978-1-4503-5029-7. doi: 10.1145/3077257.3077260. URL `http://doi.acm.org/10.1145/3077257.3077260`.

[22] R. C. Prati, G. E. A. P. A. Batista, and M. C. Monard. A survey on graphical methods for classification predictive performance evaluation. *IEEE Transactions on Knowledge and Data Engineering*, 23(11): 1601–1618, Nov 2011. doi: 10.1109/TKDE.2011.59.

[23] Chris Drummond and Robert C. Holte. Cost curves: An improved method for visualizing classifier performance. *Machine Learning*, 65(1):95–130, Oct 2006. ISSN 1573-0565. doi: 10.1007/s10994-006-8199-5. URL `https://doi.org/10.1007/s10994-006-8199-5`.

[24] Tom Fawcett. An introduction to ROC analysis. *Pattern Recogn. Lett.*, 27(8):861–874, June 2006. ISSN 0167-8655. doi: 10.1016/j.patrec.2005.10.010. URL http://dx.doi.org/10.1016/j.patrec.2005.10.010.

[25] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. A survey of methods for explaining black box models. *ACM Comput. Surv.*, 51(5):93:1–93:42, August 2018. ISSN 0360-0300. doi: 10.1145/3236009. URL http://doi.acm.org/10.1145/3236009.

[26] Mengnan Du, Ninghao Liu, and Xia Hu. Techniques for interpretable machine learning, 2018.

[27] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2017. URL https://www.R-project.org/.

[28] Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2009. ISBN 978-0-387-98140-6. URL http://ggplot2.org.

[29] E. Dimitriadou, K. Hornik, F. Leisch, D. Meyer, and A. Weingessel. *e1071: Misc Functions of the Department of Statistics (e1071), TU Wien*, 2011. URL https://CRAN.R-project.org/package=e1071.

[30] Andy Liaw, Matthew Wiener, Leo Breiman, and Adele Cutler. *randomForest: Breiman and Cutler's Random Forests for Classification and Regression*, 2018. URL https://CRAN.R-project.org/package=randomForest.

[31] Rob J Hyndman and Anne B Koehler. Another look at measures of forecast accuracy. *International Journal of Forecasting*, pages 679–688, 2006.

[32] Maxim Shcherbakov, Adriaan Brebels, N.L. Shcherbakova, Anton Tyukov, T.A. Janovsky, and V.A. Kamaev. A survey of forecast error measures. *World Applied Sciences Journal*, 24:171–176, 01 2013. doi: 10.5829/idosi.wasj.2013.24.itmies.80032.

[33] Hirotogu Akaike. *Information Theory and an Extension of the Maximum Likelihood Principle*, pages 199–213. Springer New York, New York, NY, 1973.

[34] Gideon Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, March 1978. doi: 10.1214/aos/1176344136. URL https://doi.org/10.1214/aos/1176344136.

[35] Galit Shmueli, Peter C. Bruce, Inbal Yahav, Nitin R. Patel, and Kenneth C. Lichtendahl Jr. *Data Mining for Business Analytics: Concepts, Techniques, and Applications in R*. Wiley, 2017. ISBN 1118879368.

[36] José Hernández-Orallo. ROC curves for regression. *Pattern Recognition*, 46(12):3395 – 3411, 2013.

[37] Jinbo Bi and Kristin P Bennett. Regression error characteristic curves. In *Proc. of the 20th Int. Conf. on Machine Learning*, pages 43–50, 2003.

[38] Luís Torgo. Regression error characteristic surfaces. In *KDD'05: Proc. of the 11th ACM SIGKDD*, pages 697–702, 2005.

[39] Matthew Britton. VINE: visualizing statistical interactions in black box models. *CoRR*, abs/1904.00561, 2019. URL http://arxiv.org/abs/1904.00561.

[40] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression Trees*. Wadsworth and Brooks, Monterey, CA, 1984.

[41] Igor Kononenko and Matjaz Kukar. *Machine Learning and Data Mining: Introduction to Principles and Algorithms*, chapter 6. Horwood Publishing Chichester, UK, 2007. ISBN - 10: 1-904275-21-4.

[42] Aaron Fisher, Cynthia Rudin, and Francesca Dominici. All models are wrong but many are useful: Variable importance for black-box, proprietary, or misspecified prediction models, using model class reliance, 2018.

[43] Erik Strumbelj and Igor Kononenko. A general method for visualizing and explaining black-box regression models. In *ICANNGA (2)*, volume 6594 of *Lecture Notes in Computer Science*, pages 21–30. Springer, 2011.

[44] Terence Parr, Kerem Turgutlu, Christopher Csiszar, and Jeremy Howard. Beware default random forest importances. URL https://explained.ai/rf-importance/. Accessed:20/09/2019.

[45] Jerome Friedman. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29:1217–1222, 11 2000. doi: 10.1214/aos/1013203451.

[46] Brandon Greenwell. *pdp: Partial Dependence Plots*, 2018. URL https://CRAN.R-project.org/package=pdp.

[47] Brandon M Greenwell, Bradley C Boehmke, and Andrew J Mccarthy. A simple and effective model-based variable importance measure. Technical report, 2018. URL https://arxiv.org/pdf/1805.04755.pdf.

[48] Alex Goldstein, Adam Kapelner, Justin Bleich, and Emil Pitkin. Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. *Journal of Computational and Graphical Statistics*, 24, 09 2013. doi: 10.1080/10618600.2014.907095.

[49] Alex Goldstein, Adam Kapelner, and Justin Bleich. *ICEbox: Individual Conditional Expectation Plot Toolbox*, 2017. URL https://CRAN.R-project.org/package=pdp.

[50] Daniel Apley. Visualizing the effects of predictor variables in black box supervised learning models. 12 2016.

[51] Christoph Molnar. *iml: Interpretable Machine Learning*, 2019. URL https://CRAN.R-project.org/package=iml.

[52] Jerome H. Friedman and Bogdan E. Popescu. Predictive learning via rule ensembles. *Ann. Appl. Stat.*, 2(3):916–954, 09 2008. doi: 10.1214/07-AOAS148. URL https://doi.org/10.1214/07-AOAS148.

[53] Giles Hooker. Discovering additive structure in black box functions. Technical report, 2004. URL http://faculty.bscb.cornell.edu/{~}hooker/VIN-kdd.pdf.

[54] Benjamin P. Evans, Bing Xue, and Mengjie Zhang. What's inside the black-box?: a genetic programming method for interpreting complex machine learning models. pages 1012–1020, 07 2019. ISBN 978-1-4503-6111-8. doi: 10.1145/3321707.3321726.

[55] Osbert Bastani, Carolyn Kim, and Hamsa Bastani. Interpretability via model extraction. *CoRR*, abs/1706.09773, 2017.

[56] Mark William Craven. *Extracting Comprehensible Models from Trained Neural Networks*. PhD thesis, 1996. AAI9700774.

[57] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Dino Pedreschi, Franco Turini, and Fosca Giannotti. Local rule-based explanations of black box decision systems. Technical report. URL www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

[58] Przemyslaw Biecek. *ceterisParibus: Ceteris Paribus Profiles*, 2019. URL https://CRAN.R-project.org/package=ceterisParibus. R package version 0.3.1.

[59] David Alvarez-Melis and Tommi S. Jaakkola. On the robustness of interpretability methods. *CoRR*, abs/1806.08049, 2018.

[60] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Technical report, 2018. URL https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16982.

[61] Mateusz Staniak and Przemyslaw Biecek. Explanations of model predictions with live and breakdown packages. *The R Journal*, 10, 04 2018. doi: 10.32614/RJ-2018-072.

[62] Lloyd S Shapley. A value for n-person games. In Harold W. Kuhn and Albert W. Tucker, editors, *Contributions to the Theory of Games II*, pages 307–317. Princeton University Press, Princeton, 1953.

[63] Scott M. Lundberg, Gabriel G. Erion, and Su-In Lee. Consistent individualized feature attribution for tree ensembles. *CoRR*, abs/1802.03888, 2018.

[64] Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *NIPS*, pages 4768–4777, 2017.

[65] Anupam Datta, Shayak Sen, and Yair Zick. Algorithmic transparency via Quantitative Input Influence: Theory and experiments with learning systems. In *Proceedings of the 2016 IEEE Symposium on Security and Privacy (SP)*, pages 598–617. IEEE, May 2016. ISBN 978-1-5090-0824-7. URL http://dx.doi.org/10.1109/sp.2016.42.

[66] Giuseppe Casalicchio, Christoph Molnar, and Bernd Bischl. Visualizing the feature importance for black box models. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 11051 LNAI:655–670, 2019. ISSN 16113349. doi: 10.1007/978-3-030-10925-7_40.

[67] Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *SSRN Electronic Journal*, 11 2017. doi: 10.2139/ssrn.3063289.

[68] Josua Krause, Adam Perer, and Kenney Ng. Interacting with predictions. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems - CHI '16*, pages 5686–5697, May 2016. doi: 10.1145/2858036.2858529. URL `http://dl.acm.org/citation.cfm?doid=2858036.2858529`.

[69] John W. Tukey. *Exploratory Data Analysis*. Addison-Wesley, 1977.

[70] A. N. Kolmogorov. Sulla determinazione empirica di una legge di distribuzione. *Giornale dell'Istituto Italiano degli Attuari*, 4:83–91, 1933.

[71] N. Smirnov. Table for estimating the goodness of fit of empirical distributions. *Ann. Math. Statist.*, 19(2):279–281, 06 1948. doi: 10.1214/aoms/1177730256. URL `https://doi.org/10.1214/aoms/1177730256`.

[72] T. W. Anderson and D. A. Darling. Asymptotic theory of certain goodness of fit criteria based on stochastic processes. *The Annals of Mathematical Statistics*, 23(2):193–212, 06 1952. doi: 10.1214/aoms/1177729437. URL `https://doi.org/10.1214/aoms/1177729437`.

[73] Two sample kolmogorov-smirnov table. URL `http://www.real-statistics.com/statistics-tables/two-sample-kolmogorov-smirnov-table/`. Accessed:7/08/2019.

[74] Sonja Engmann and Denis Cousineau. Comparing distributions: the two-sample Anderson–Darling test as an alternative to the Kolmogorov–Smirnov test. *Journal of Applied Quantitative Methods*, 6: 1–17, 09 2011.

[75] Fritz Scholz and Angie Zhu. *kSamples: K-Sample Rank Tests and their Combinations*, 2019. URL `https://CRAN.R-project.org/package=kSamples`.

[76] F. W. Scholz and M. A. Stephens. K-sample Anderson-Darling tests. *Journal of the American Statistical Association*, 82(399):918–924, 1987. doi: 10.1080/01621459.1987.10478517.

[77] Alfred Inselberg. The plane with parallel coordinates. *The Visual Computer*, 1(2):69–91, 1985. URL `http://dblp.uni-trier.de/db/journals/vc/vc1.html#Inselberg85`.

[78] Max Kuhn. *caret: Classification and Regression Training*, 2019. URL `https://CRAN.R-project.org/package=caret`.

[79] Justin Alger. *Large Marine Protected Areas and Ocean Resilience: Stakeholder Conflict in Pelagic Seas*, pages 168–183. Cambridge University Press, 2019. doi: 10.1017/9781108502238.011.

[80] Chris Smyth and Quentin Hanich. *Large Scale Marine Protected Areas: Current status and consideration of socio-economic dimensions*. Australian National Centre for Ocean Resources and Security (ANCORS), 2019.

[81] The International Union for Conservation of Nature. Large scale marine protected areas. URL `https://www.iucn.org/commissions/world-commission-protected-areas/our-work/large-scale-marine-protected-areas`. Accessed:19/09/2019.

[82] Dainel Wagner, Aulani Wihlem, Alan Friedlander, Andrew Skeat, Anne Sheppard, Brian Bowen, Carlos Gaymar, Gustavo Martin, Ian Wright, Jason Philibotte, John Parks, Jolene Bosanquet, Kahoane Aiona, Joseph brider, Kim Morishige, Liz Wright-Koteka, Nai'a Lewis, Noeline Brownie, Randall Kosaki, and Zeenatul Basher. Big Ocean - A shared research agenda for Large-Scale Marine Protected Areas. 02 2013.

[83] Kristina Boerder, Bethan O'Leary, Chris McOwen, Elizabeth Madin, Douglas McCauley, Caroline Jablonicky, Luis Torgo, Manuel Dureuil, Derek P. Tittensor, and Boris Worm. Interactions between large marine protected areas and global fishing fleets. *(under review)*, 2019.

[84] Fao fisheries & aquaculture - effects, benefits and costs of mpas. URL `http://www.fao.org/fishery/topic/16201/en`. Accessed:19/09/2019.

[85] Luis Torgo. *performanceEstimation: An Infra-Structure for Performance Estimation of Predictive Models*, 2016. URL `https://CRAN.R-project.org/package=performanceEstimation`.

[86] Stephen Milborrow. *earth: Multivariate Adaptive Regression Splines*, 2019. URL `https://CRAN.R-project.org/package=earth`.

[87] Bruce Ratner. *Statistical and Machine-Learning Data Mining: Techniques for Better Predictive Modeling and Analysis of Big Data*. Taylor & Francis, 3 edition, 2017.

[88] Global fishing watch. URL `https://globalfishingwatch.org`. Accessed:10/09/2019.

[89] Daniel Apley. *ALEPlot: Accumulated Local Effects (ALE) Plots and Partial Dependence (PD) Plots*, 2018. URL `https://CRAN.R-project.org/package=ICEbox`.

[90] Taiyun Wei, Viliam Simko, Michael Levy Levy, Yan Xie, Yihuiand Jin, and Jeff Zemla. *corrplot: Visualization of a Correlation Matrix*, 2017. URL `https://CRAN.R-project.org/package=corrplot`.

Appendix A

Trivariate EDPs: Two Examples

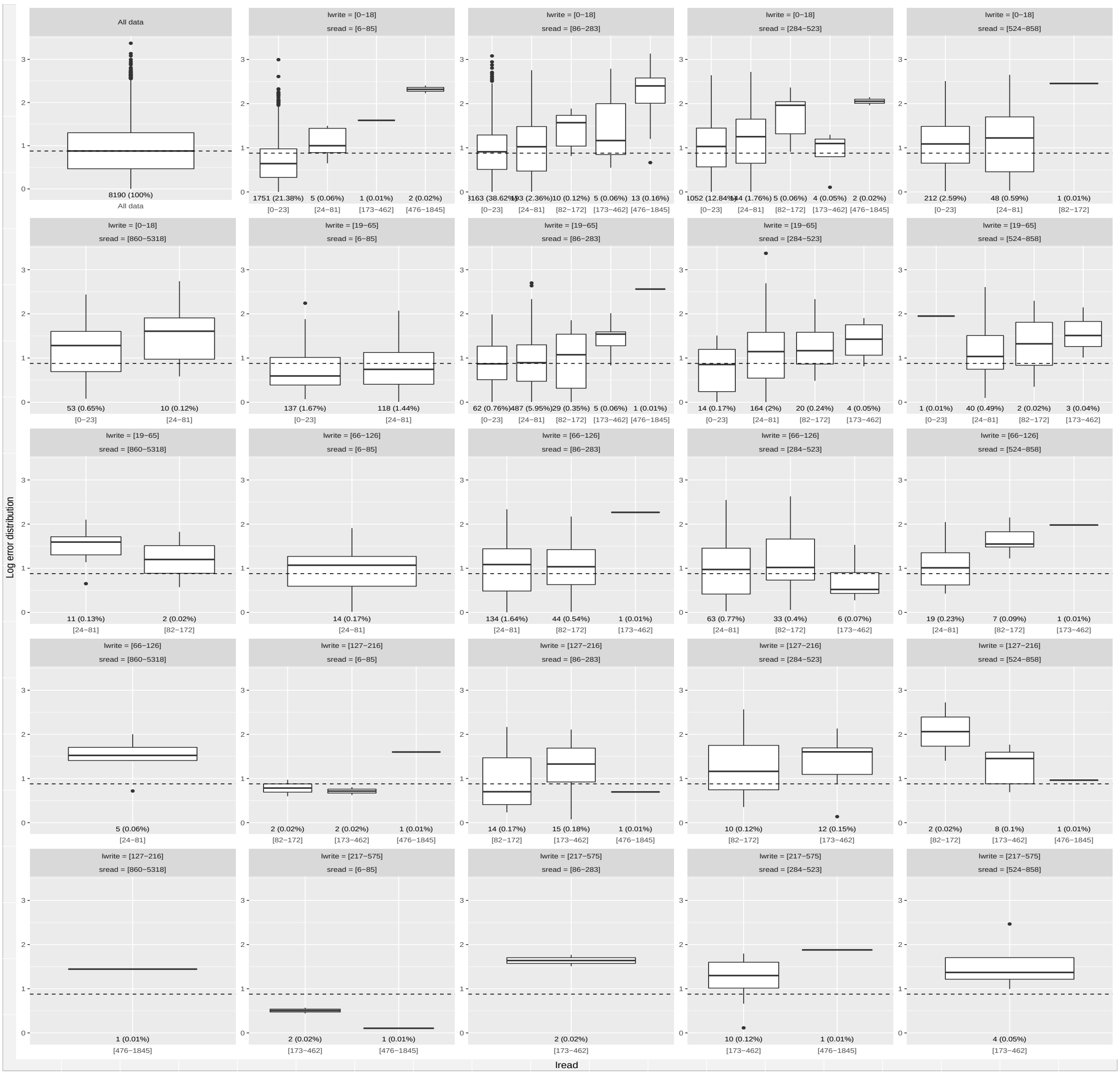


Figure A.1: Trivariate EDP of a RF for data set *cpuSM* for interactions between *lwrite*, *sread* and *lread*in *condensed* mode.

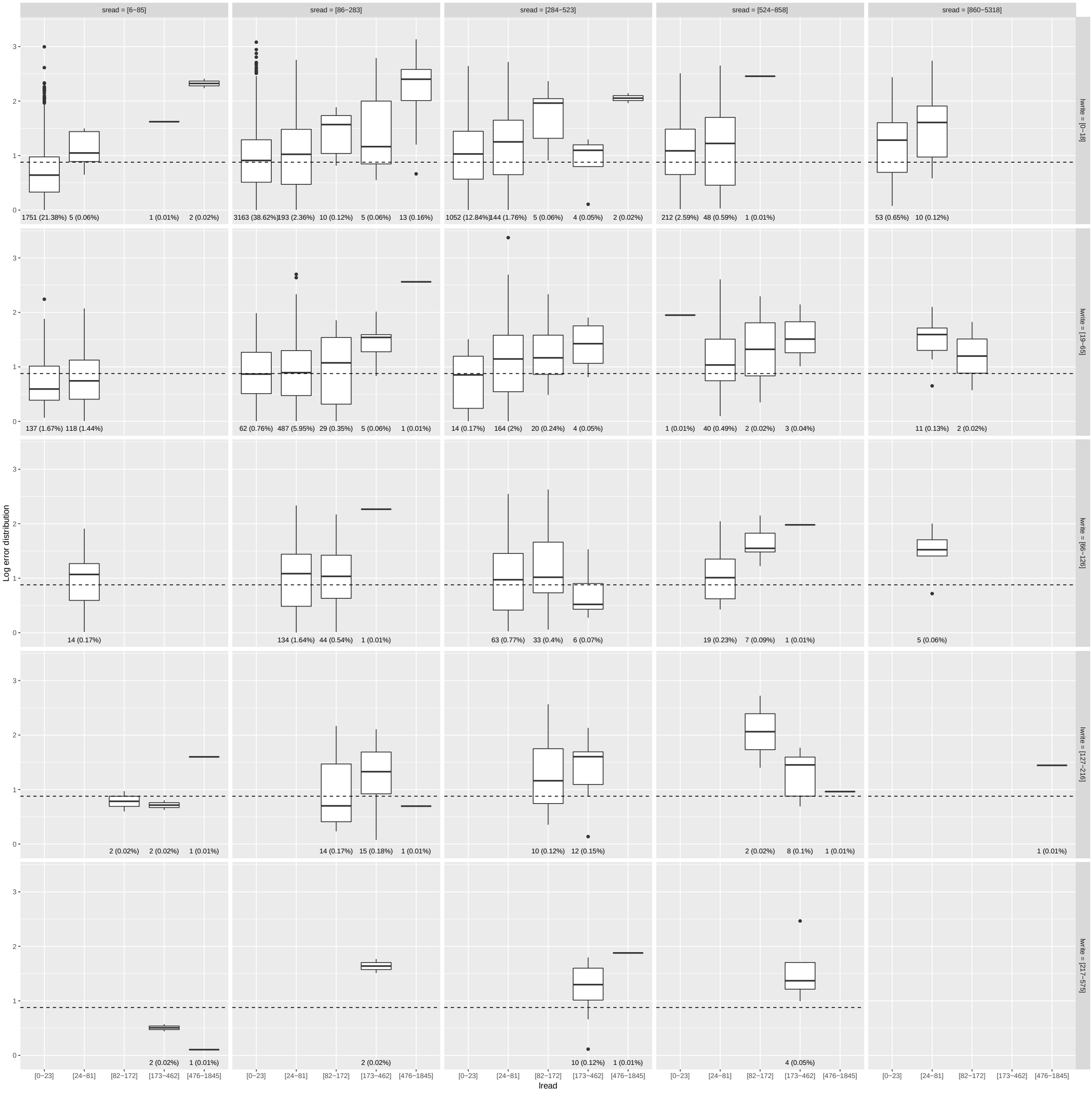


Figure A.2: Trivariate EDP of a RF for data set *cpuSM* for interactions between *lwrite*, *sread* and *lread* in *grid* mode.

Appendix B

LSMPA Support Analysis

B.1 LSMPAs Characteristics

MPA	Age	Size	GDP	Shape	%MPA bordering High Seas	%No-take	Number protected status categories	Enforcement	%Buffer zone in High Seas	Number of zones
Argo-Rowley Terrace MP	5	146,003	1,275,000	74.9	19	0	3	1	18.0	4
Chagos MPA	7	639,661	2,739,988	200.4	81.9	100	1	1	64.7	1
Coral Sea MP	5	989,924	1,275,000	177.4	0	0	4	1	2.2	26
Galapagos MR	19	138,000	98.990	79.3	0	0	0	2	41.0	1
Great Barrier Reef MP	42	348,700	1,275,000	48.7	0	33	8	3	0.3	325
Lord Howe MP	5	110,139	1,275,000	54	6.2	0	5	1	41.3	7
Macquarie Island MP	18	162,000	1,275,000	90.7	36.7	36	2	2	48.2	3
Marianas Trench MNM	8	246,608	18,302,874	52.2	4	17	2	1	34.6	3
Natural Park of the Coral Sea	3	1,291,000	2,449,500	182.3	17.5	0	0	1	15.6	0
Norfolk MP	5	188,444	1,275,000	74.3	7.5	0	2	1	33.7	2
Pacific Remote Islands MNM	18	1,269,094	18,302,874	143.8	63.2	100	2	1	63.8	10
Papahanaumokuakea MNM	17	362,061	18,302,874	104.8	0	100	3	3	43.9	11
Phoenix Islands PA	11	408,250	163	160.3	0	994	2	1	36.8	2

Table B.1: Characteristics of Large-Scale Marine Protected Areas (LSMPAs) as of 2015 included in the study [83].
Units: *Age*: years, *Size*:km^2, *GDP*: million$, *Shape*: km

B.2 Performance Analysis

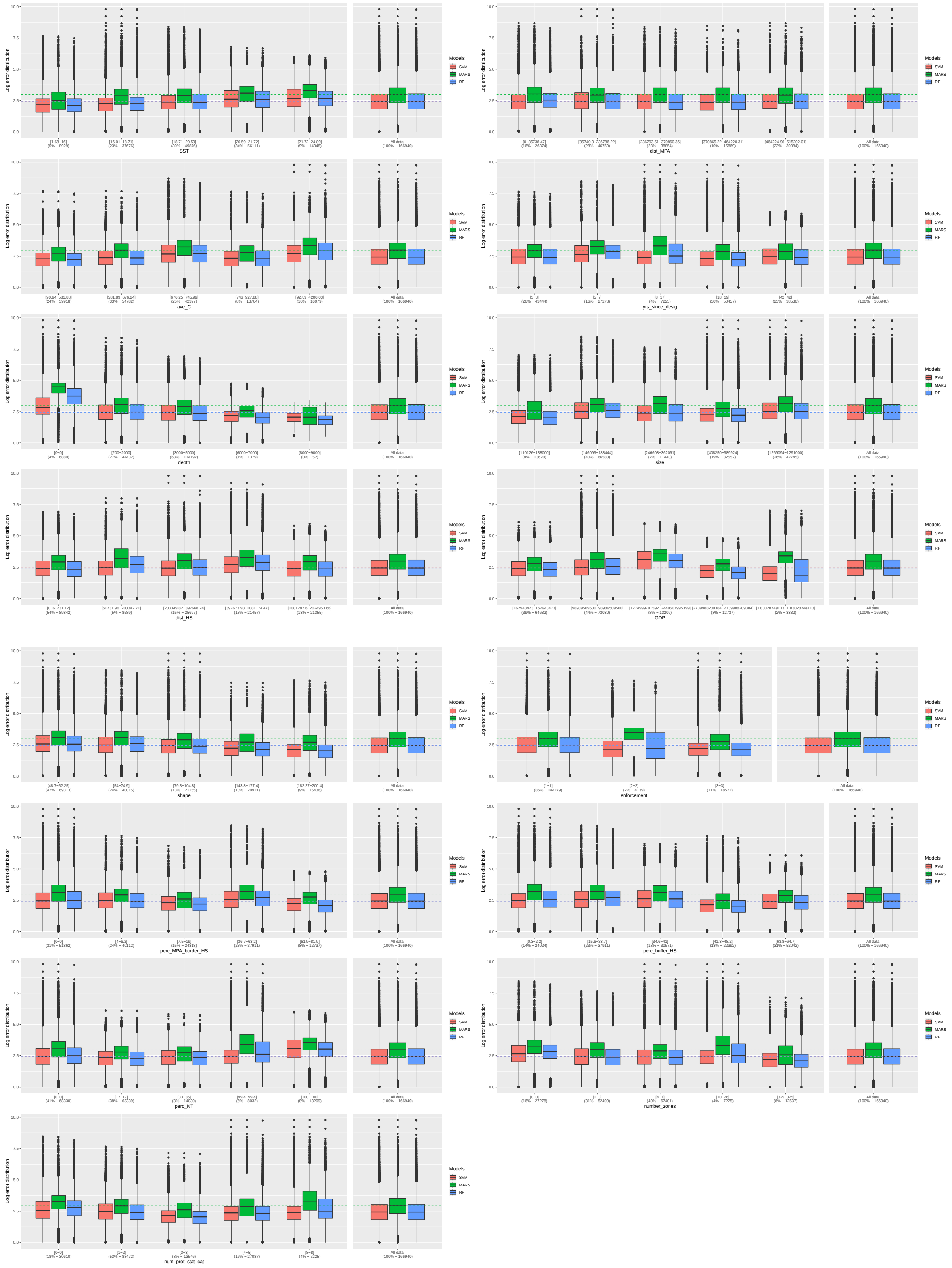
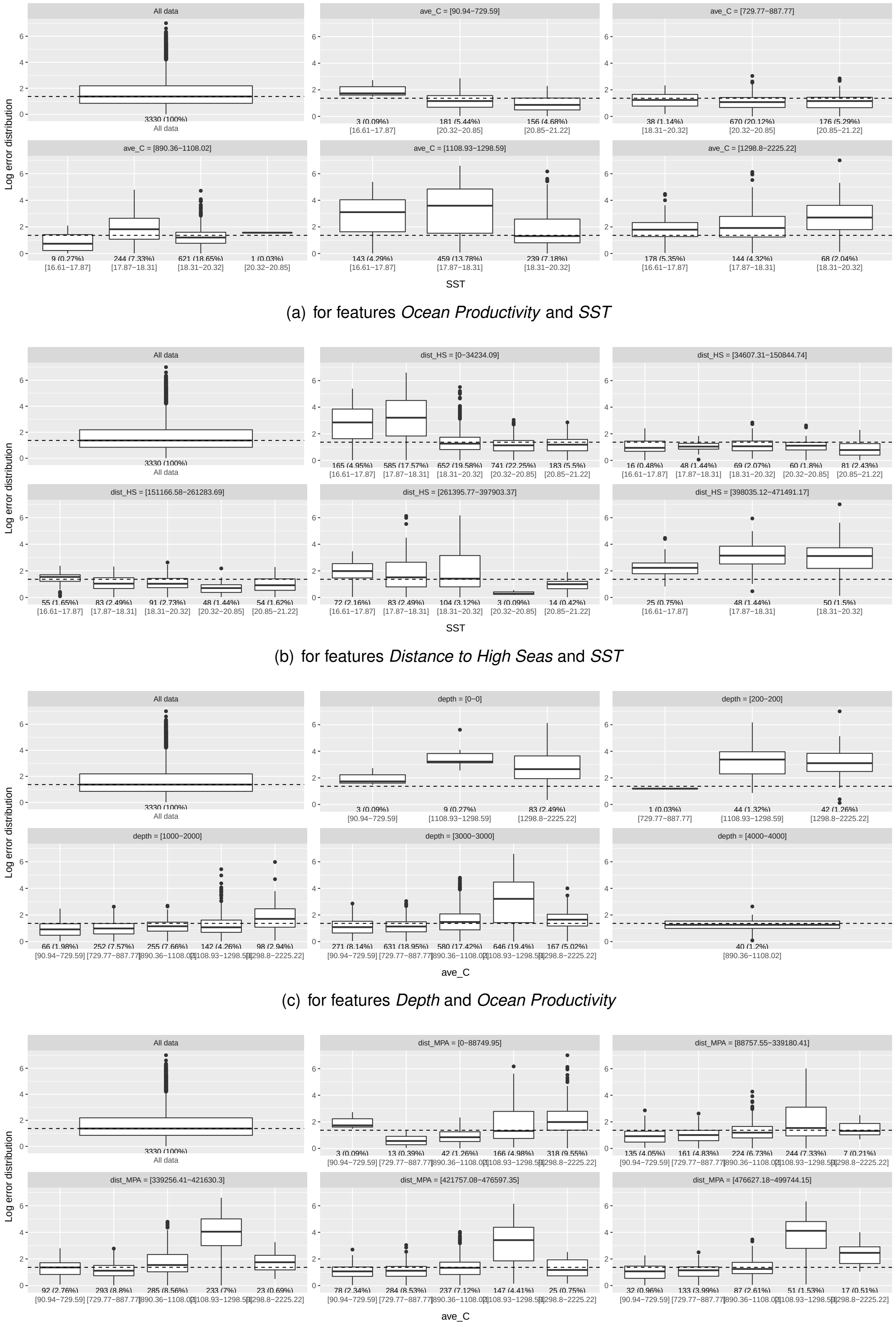


Figure B.1: Multiple model Error Dependence Plot for All MPAs, in respect to each predictor variable values



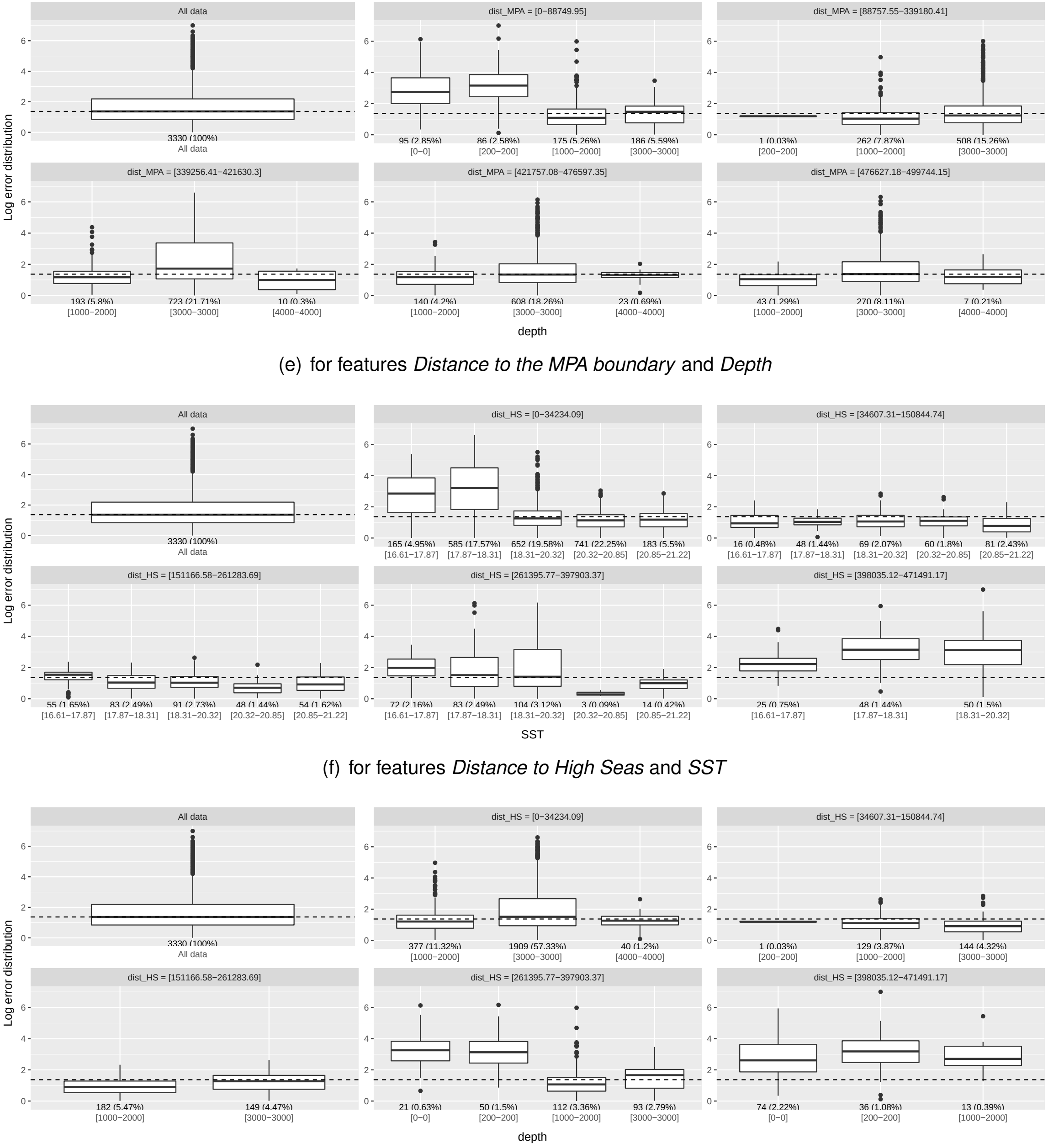
(a) for features *Ocean Productivity* and *SST*

(b) for features *Distance to High Seas* and *SST*

(c) for features *Depth* and *Ocean Productivity*

(d) for features *Distance to the MPA boundary* and *Ocean Productivity*

Figure B.2: Bivariate EDPs for the Random Forest that models the fishing effort in Galapagos MR



(e) for features *Distance to the MPA boundary* and *Depth*

(f) for features *Distance to High Seas* and *SST*

(g) for features *Distance to High Seas* and *Depth*

Figure B.3: Bivariate EDPs for the Random Forest that models the fishing effort in Galapagos MR (cont.)

B.3 Correlation Plots

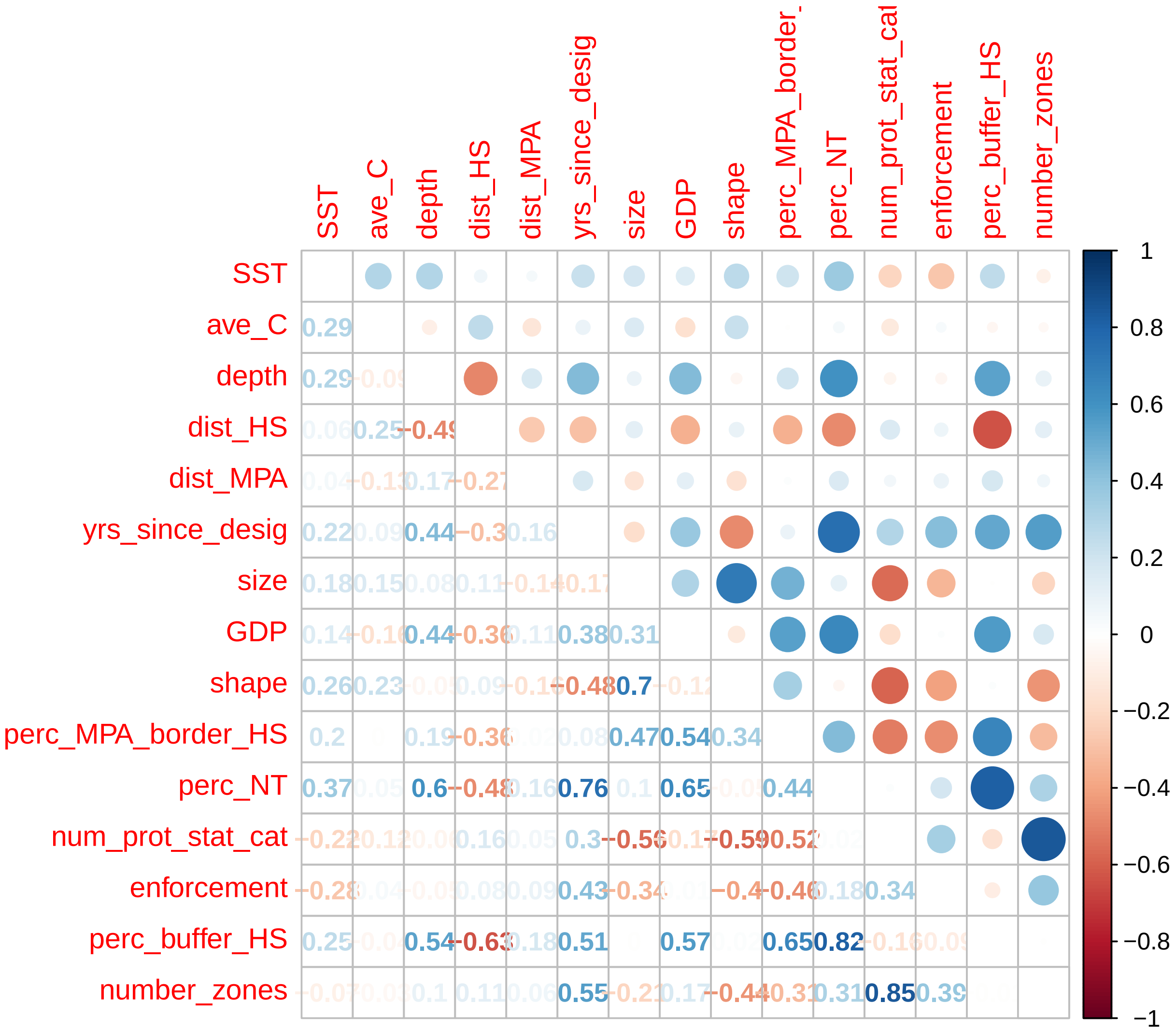
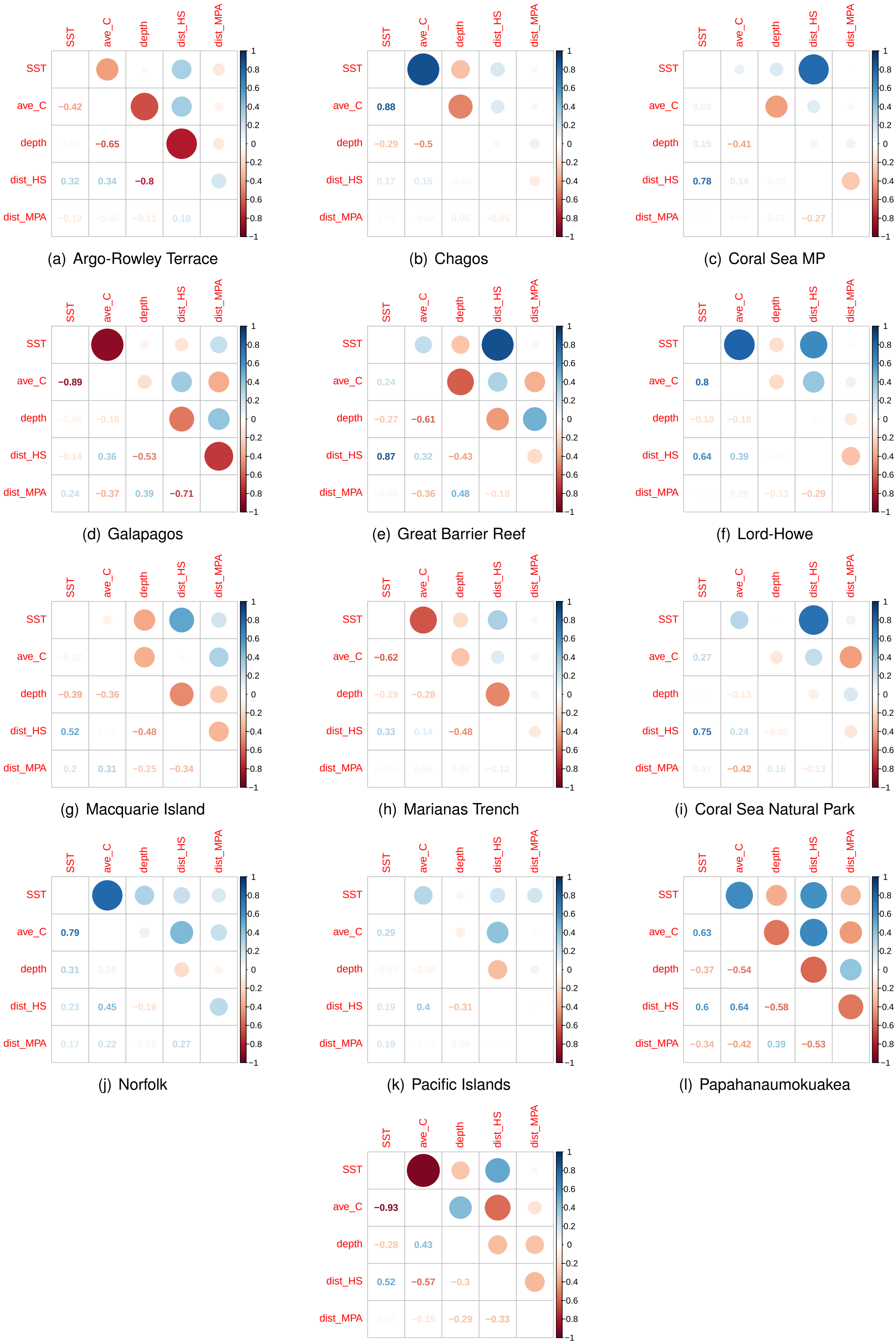


Figure B.4: Correlation between predictor for All LSMPAs. Calculated using library *corrplot* [90], with Spearman coefficient. The bigger the circle, the stronger the correlation between the two features. Blue colour represents a positive correlation while red colour represents a negative correlation.



(a) Argo-Rowley Terrace

(b) Chagos

(c) Coral Sea MP

(d) Galapagos

(e) Great Barrier Reef

(f) Lord-Howe

(g) Macquarie Island

(h) Marianas Trench

(i) Coral Sea Natural Park

(j) Norfolk

(k) Pacific Islands

(l) Papahanaumokuakea

(m) Phoenix Islands

Figure B.5: Correlation matrix plot between predictors for each MPA. Calculated using library *corrplot* [90]