The York Computer Science Department

Model interrogation: Surrogate models informed by counterfactual examples for model interpretability

MSC dissertation

Alex Varden

Table of Contents

[1 Introduction 3](#_Toc130926557)

[1.1 Focus of Research 3](#_Toc130926558)

[1.2 Questions 3](#_Toc130926559)

[1.2.1 Question 1 3](#_Toc130926560)

[1.2.2 Question 2 3](#_Toc130926561)

[1.2.3 Question 3 3](#_Toc130926562)

[2 Literature Review 4](#_Toc130926563)

[2.1 Black boxes 4](#_Toc130926564)

[2.1.1 Neural Networks 4](#_Toc130926565)

[2.1.2 Random Forests 4](#_Toc130926566)

[2.2 Surrogate Models 4](#_Toc130926567)

[2.2.1 LIME 5](#_Toc130926568)

[2.2.2 Decision trees as surrogate models 5](#_Toc130926569)

[2.3 Counterfactuals 6](#_Toc130926570)

[3 Theory 8](#_Toc130926571)

[4 Method 9](#_Toc130926572)

[4.1 Generating counterfactuals 9](#_Toc130926573)

[4.1.1 generating perturbations 9](#_Toc130926574)

[4.1.2 localisation 9](#_Toc130926575)

[4.1.3 multiclass classification 10](#_Toc130926576)

[4.1.4 Regression 10](#_Toc130926577)

[4.1.5 Surrogate Model 10](#_Toc130926578)

[4.2 Experimental Method 12](#_Toc130926579)

[4.2.1 Models under analysis 12](#_Toc130926580)

[4.3 Measurement 13](#_Toc130926581)

[4.3.1 Classification 13](#_Toc130926582)

[4.3.2 Regression 13](#_Toc130926583)

[4.3.3 Complexity 13](#_Toc130926584)

[4.4 Experiment 1 – Global Fidelity 13](#_Toc130926585)

[4.5 Experiment 2 – Local fidelity 14](#_Toc130926586)

[4.6 Experiment 3 – Counterfactual 14](#_Toc130926587)

[5 Results 15](#_Toc130926588)

[5.1 Hardware 15](#_Toc130926589)

[5.2 Models Under Analysis 15](#_Toc130926590)

[5.3 Experiment 1 - Global Surrogate Modal 16](#_Toc130926591)

[5.3.1 Time analysis 16](#_Toc130926592)

[5.3.2 Performance 16](#_Toc130926593)

[Bibliography 17](#_Toc130926594)

# Introduction

## Focus of Research

Two key metrics of any predictive model are accuracy and interpretability, the efficacy of black-boxed techniques such as neural networks have encouraged researchers to prefer the former over the latter. which are being used to inform high stakes decision domains such criminal recidivism prediction [1], and healthcare domains [2]. Backboxes are notoriously hard to debug, “Making the right decision for the wrong reasons” [26] leading often to models that under perform in the real world [6], [27].

Interpretability techniques have been criticized for being unfaithful to how it approximates the backbox’s decision [12]. Until inherently interpretable models become sufficiently accurate, we must attempt to overcome their limitations of backboxes and accept Box’s maxim "All models are wrong, but some are useful."

I have identified two key techniques for model interpretability, surrogate models [6],[8],[13] & counterfactual examples [18], [20] which Mothilal et al describes as “Different means to the same end” [16]. I will explore Wachter’s [18] counterfactual optimization applied to improving the accuracy of a surrogate model.

I aim to develop a model agnostic post-hoc technique which can be used to improve the fidelity of a surrogate model. I will explore both global and localised versions of this method to ascertain if the technique is more effective with a reduced scope.

## Questions

### Question 1

Can training a surrogate model on counterfactual examples improve the fidelity/accuracy to that of the model under analysis over random sampling?

### Question 2

How does localisation improve the quality of the explanations given by the surrogate model and what methods can we use to achieve this?

### Question 3

How does altering proximity and diversity effect the accuracy or complexity of the surrogate model?

# Literature Review

## Black boxes

### Neural Networks

An artificial neural network (ANN) is a set of nodes that are connected in various topologies to create a network. Each node is a perceptron which takes a weighted sum [53] of its inputs and preforms a non-linear activation function to output a value [54]. These nodes are organised into layers which are often densely connected.

These networks are exposed to training data and their error is calculated in relation to a desired output. The error is minimised through gradient descent which calculates the slope of the error function in an N dimensional plane. The activation function ensures that values form previous inputs are continuous which ensures the gradient is smooth so that it may find a local minima [56]. The network updates its weights through each layer of the network through a process called backpropagation [54].

ANNs can represent complex non-linear relationships [53] in the data in fact Hornik [54] showed that a multi-layer perceptron is a universal function approximator capable of mapping any given input with any output. However due to the complex nature of the networks they are often regarded as backboxes and any meaningful information about how a particular networks arrives at a decision is inaccessible.

### Random Forests

An individual decision tree is an acyclic directed tree which represents a binary decision at each branch. These trees are generated through continuously splitting data either by entropy gain or minimising GINI impurity. These simple structures are inherently interpretable and are able classify non-linear relationships [7]. However, trees have been shown to be unstable in that minor changes to the input can have major effect on the structure of the tree. They also suffer from not being very accurate.

Random forests [2001] utilises this instability to improve performance by generating multiple decision trees with varied parameters. The ensemble makes predictions based on votes cast by individual trees to give a probability.

Interpretable machine learning seeks to approximate the structure of black boxes. So that we may interrogate its decision process. Molnar’s book [11] provides an excellent overview of the subject, whilst making good use of citation of popular research in the field.

## Surrogate Models

Knowledge distillation [9] [10], attempts to replace complex model (teacher) with techniques that are inheritably interpretable (student), such as a linear model or decision tree. They train a student model on teacher’s prediction labels. This often comes at the expense of a lower prediction accuracy [10].

Surrogate models differ in their desiderata. The student model is treated as a cartographic representation of its teacher. This allows developers to map the structure of its decision boundary without sacrificing accuracy of the final model.

### LIME

A widely cited techniques for achieving this is Locally Interpretable Model-agnostic Explanations (LIME) [6]. This post-hoc surrogate model provides locally faithful feature importance weights for a model’s decision. This is achieved through training a surrogate model on generated perturbations with a high proximity to the explanatory value. As can be seen in [equation 1] the technique balances the ‘Fidelity-Interpretability trade-off’ [6] by selecting the potential explanation *g* from all possible Explanations *G by* minimising the accuracy loss *L* with perturbations within the kernel radius alongside the complexity . This results in an explanation which is both locally accurate and simple enough to understand.

Equation. 1

Lime achieves locality through generating random samples that have a similar mean and gaussian distribution as the global feature space. It uses a gaussian radial basis function (RBF) to plot distance in hyperdimensional to the given example. It then uses these distances as a weighted input to its surrogate model [66].

Equation. 2

RBF [equation 2] is a generalisation of Euclidian distance which applies a scaling factor known as the kernel length. Molnar [11] warns about the practical pitfalls in defining the size of lime’s kernel. It’s local explanations can fail to grasp nuances of the model due to their limited scope. In contrast Laugel et al [66] argues that due to lime drawing samples from the global feature spaces it can miss locally important features in it’s explanations in favour of global important.

Laugel [66] discus different methods for establishing an area to localised perturbations, they compare limes [6] weighted methods to their own ‘local surrogate’ which samples from a hypersphere whose radius grows until it encounters a counterfactual example. Their test shows an improvement in local fidelity of each model when compared to lime with default kernel and a kernel length of 0.5.

[] Conducted an extensive literature review of previous LIME evaluations. They Identify a gap in lime being applied to tabular data. The paper takes a pedagogical approach to explaining their method in the hope they can set a standard of how future post-hoc evaluations of explainable AI methods could be conducted. However, the generalisability of their method could have been improved by the inclusion of multiple datasets as is popular in machine learning evaluation as discussed by Demšar [38] and demonstrated by [8,6]. Their mixed methods approach echoed that of the original study conducted by [6] evaluating the method quantitively through experimental comparison of performance and then qualitatively evaluating the perceived usefulness of the output with interviews.

### Decision trees as surrogate models

Decision trees have frequently been used as surrogate models [8,13] they are naturally human interpretable. however, they do become hard to be interpreted by a lay person the larger the get and so it stands to reason that the most useful trees are as short as they can be. Unfortunately generating an optimal binary classification tree has been shown to be an NP complete problem [15].

Bennett[] uses global tree optimization convert a CART Decision tree into a collection of disjunctive linear inequalities, they apply extreme point tabu search to find a globally optimal multivariant tree of a given depth. The draw back of a multivariant tree (also known as an oblique tree) is that they trade tree depth with complexity at each node. this results in less interpretability.

Izza[] challenges the widely accepted use of decision trees as an interpretable model and claims they are In fact not interpretable due to their being a large number of redundant paths. They show this through an analysis of decision trees generated on UCI datasets. They judge a path to be redundant if one level of the tree could be inferred from the others.

Bastani [13] uses Exact greedy decision tree as a surrogate model, they demonstrate their method’s model agnosticism by testing on a range of predictive model types; regression, classification and reinforcement learning. They show a consistent gain in accuracy across all models. The paper purports to show that trees are effective as surrogate models however their experiment compares two trees against each other which doesn’t validate their initial claim. They do however give examples of how trees can be used in practice. I attribute the paper brevity to the fact that it was presented as a poster for a Workshop.

Saadallah [8] show surrogate model accuracy can be improved through a sampling technique which uses Query-By-Committee (QBC) it identifies contentious instances using a variance-based measure of disagreement. Saadallah [8] refers to Bastani [13] axiom, that given infinite samples a greedy decision tree student model will converge on is teacher. They conclude that “in practice, we are limited by the amount of data available to train the surrogate model” [8]. They aim their research towards finding quality samples over quantity. In my research we will attempt to overcome the limitation of quantity by providing a means for generating more training data from existing data whilst also improving the quality.

Demšar [38] discusses good statistical practices in machine learning research, They would describe [8, 6] as typical of recent machine learning research in that it takes a method, applies it to a range of datasets and compares its performance across them. My own research will take a similar methodological structure. [8] uses a Wilcoxon t-test to draw conclusions whereas Bastani [13] counts wins, and losses. Both are inline with Demšar recommendation; however the Wilcoxon t-test is the more powerful technique [38].

## Counterfactuals

Counterfactuals (CF) are hypothetical instances which minimize the difference that give an opposing classification. Mittelstadt [17] reasons that everyday human explanations are naturally contrastive due to the “cognitive complexity of non-contrastive explanations” and so we understand causal relationships through imagining how it would change given a different scenario.

Wachter [18] developed a novel optimization to compute CF explanations. their optimization incrementally increases the distance from a given example until it finds a contrastive example. This could be thought of as identifying how close a given example is to the decision boundary.

The paper spends most of its text situating CFs within philosophical and legal framework of GDPR’s “Right to an explanation” The papers main contribution was its optimisation function [Equation 3]

Equation. 3

One drawback highlighted by Wachter[18] is that hypothetical examples given may not be a valid input for the feature space or within the control of the person receiving the explanation to change, for this reason it is advantageous to be able to control the diversity and proximity of this optimization, a contribution which was made by Mothilal [19] builds upon Watcher’s work with a more practical approach. Developing a framework, DiCE [21] which parameterised diversity and proximity. Diversity is a measure of how many features can be altered and proximity is by how much.

The *Rashmond effect* [11] is a phenomenon in which multiple different perspectives are valid, in the context of CFs there are equally explanative counter examples for any given instance.

Dice & Watchers method depends on access to a neural network’s gradients, this makes it model dependent. There are however extensions to Dice [21] which allow it to perform its optimization through a genetic algorithm allowing it to be model agnostic. Selecting meaningful counter examples is the subject of much research [19,20,22] however for my research my research does not need them to be explanative but does value their proximity. [22] draws comparisons between counterfactual explanations and adversarial examples, in that they differ only in their desiderata. An observation also made by Wachter [18].

Mothilal [16] shows that feature attribution techniques such LIME and SHAP do not agree with feature importance rankings of counterfactuals generated by Wachter [18] and Dice [21] method. They generate feature attribution scores from counterfactual methods, and compare against SHAP and Lime, whilst SHAP and Lime attribution are consistent with each other they find that they are not consistent with those generated by CF methods.

They measure necessity and sufficiency. A feature is necessary if varying it whilst fixing other features would result in a counterfactual, whereas a sufficient feature would be one that when fixed, varying other features does not result in a counterfactual. Their results found that Lime & SHAPs top features are not very necessary and found they are overwhelmingly insufficient in that none of the important features could prevent a flip in classification.

Their findings show that lower ranked features identified by LIME or SHAP can have a large effect on the outcome of a classification which may lead to over valuing their output. Integrating counterfactuals into a surrogate model such as Lime could possibly lead to a reconciliation of their feature attribution.

# Theory

A surrogate model can be likened to an interviewer interrogating an uncooperative subject trying to tease out details by asking questions and recording their responses.

The interrogator can ask as many questions as they would like, and the subject will give honest answers (from their perspective). The job of an effective interviewer is to select the questions that reveal the most about the subject.

Random perturbation techniques such as LIME ask a lot of questions however their stochastic nature doesn’t allow them to direct the investigation towards the edges of the decision boundary.

CF examples generated through Wachter’s optimization [18] minimize their proximity to the decision boundary, I posit that like Saadallah [8] identifying contentious instances, these CFs can provide quality questions for a surrogate model to learn from.

It logically follows that a counterfactual optimization applied for two generations should result examples on either side of a decision boundary with a minimized proximity to each other. Analogues to how support vector machines represent a decision boundary [49] by only paying attention to the closest vectors to it. (See figure 1)

A picture containing shape

Description automatically generated

Figure : 2 generations of CF examples and their relation to the decision boundary

# Method

I have identified two models which may complement one another. Through abductive reasoning I have posited a theory which, in this research I hope to test through deductive reasoning [28].

In the following section I will first describe the surrogate model technique that I have developed and give justifications for key decisions, I will then go on to describe how I will evaluate this technique quantitively through experiment, if the method shows to be effective within the bounds of the experiment, then we may infer that its results generalize to other similar settings.

## Generating counterfactuals

The method uses DiCE [21] which provides an open-source implantation of watchers [18] optimization to generate counterfactuals. Dice and watcher where both designed to use the weights of a neural network to perform its optimization. One criteria for this method was for it to remain model agnostic and thankfully dice provides the availability to run the optimisation through a genetic algorithm which does not require the weights as input. It is a requirement that one must specify the opposite class which you wish to convert the given example too.

Dice parameters allow control for :

**Proximity:**

**Diversity:**

**Sparsity:**

### generating perturbations

Traditional counterfactuals explanations are designed to be an explanation of a single input value however, we wish to generate a large volume of counterfactuals. Selecting the input values will dictate where in the feature space the counterfactual will begin and thus which area they will explain. Similar to Lime’s [6] method we draw a random hypothetical examples which have a similar mean and distribution as the global feature space.

For continuous data I was able use `numpy.random.normal()` [67] which generates a random series with a given mean and standard deviation. For categorical features I was able to maintain the frequency of each value with `np.random.choice` [68]. Once these examples are generated, we can feed them to our model under analysis and store its prediction. This forms the training data for our initial training data for our surrogate model.

### localisation

Localisation is achieved through applying a nearest neighbour (NN) algorithm to the global perturbations. NN utilizes k means clustering to represent all features in a hyperdimensional space and computes **Minkowski distance between a given example and the other instance.** One method for selecting NN is a creating a hyper dimensional sphere with a given radius *R* as proposed by Laugel *[66].* This method selects the purest subset which are the most proximal to each other. The other method is specifying how large you would like the subset to be, and it will return k nearest neighbours. This results in a subset which is a fixed size however may include instances which may not be a natural fit for each other.

Because the nearest neighbour algorithm is geometric in nature It is important to normalise all values so that it can have a “apples for apples” comparison when representing it spatially. For continuous values this is easily solved by using Minmax scaling to force values to lie between 0 and 1, however a problem presents itself with one hot encoded categorical value as any change in a value is 1 which represents a sphere that engulfs the entire feature space. For this reason, it makes sense to use K closest for datasets which have many categorical values such as the adult dataset.

once the nearest neighbours have been identified they become the seed for a further batch of perturbations with the same mean and distribution as the local neighbourhood.

### multiclass classification

In a multi-class classification problem, there are multiple opposite values to a given example. Without a good priori way to know the closest opposite class the simplest method I have found is to generate all other classes present in the sample, this however this results in slightly less than n2 comparisons. To constrain this, I randomly sample a percentage relative to the number of classes to keep n constant.

Text

Description automatically generated

Figure 2 : Multiclass counterfactual generation (implmentation details left out for clarity)

### Regression

Counterfactuals in the context of a regression problems require specifying a desired range which you wish model to output. As with classification we want these desired range to represent a substantive difference in prediction. To ensure this I require a counterfactual to be a part of a different quantile range, this forces the counterfactual target to be sufficiently far from its original value. By treating the quantile as a class in this way, I can use the same logic as multiclass counterfactuals and generate one for each quantile range.

### Surrogate Model

Once the training data for my surrogate model has been generated in the form of counterfactual examples, I will train a model to predict the output of the model under analysis.

**Decision tree**

A decision is an acyclic directed tree which represents a binary decision at each branch. These trees are generated through continuously splitting data either by entropy gain or minimising GINI impurity.

CART Decision trees use Gini impurity [15] at each split to determine.

Their greedy nature results in dominant attributes being closer to the top. This makes them the suited for showing feature importance as highest contributing factors are closer to the top of the tree [13].

Discretize data !

Linear Model

## Experimental Method

Having established a method for generating counterfactual-based surrogate-models this section focuses on the method I will use to quantitively test its efficacy.

This research seeks to understand an ontologically deterministic process and so a post-positivist epistemology is appropriate [29]. Post-positivism builds on positive empiricism however accepts that knowledge is antifoundational [29] for this reason through falsification I will attempt to reject my null hypothesises (H0) through quantitively experiment.

### Models under analysis

Blockeel advocate for “varying the conditions under which [machine learning research] experiments are run” [24] to ensure results are generalizable UCI machine learning repository [36] hosts a vast range dataset used for popular machine learning examples. I have selected 5 of the most popular datasets (as rated by clicks on UCI) As my aim is to demonstrate the model agnosticism of my technique, I have ensured that there are mixture of regression and classification problem types. Using datasets that are open and publishing my code, I hope to increase the reproducibility of my findings [30].

Table

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| # | Ref | Name | attributes | Instances | Problem type |
| 1 | iris | Iris Data Set [31] | 4 | 150 | Multivariate Classification |
| 2 | adult | Adult Data Set [32] | 14 | 48842 | Classification |
| 3 | bean | Dry Bean Dataset [33] | 17 | 13611 | Multivariate  Classification |
| 4 | wine | Wine Quality Data Set [34] | 12 | 4898 | Regression |
| 5 | bike | Bike Sharing Dataset [35] | 16 | 17389 | Regression |

For the sake of this experiment the performance of the models being interpreted aren’t a primary concern. However, to ensure that they are representative of models found ‘in the wild’ I have applied some techniques to improve the predictive performance of the models.

I have used a popular machine learning package for python scikit learn. Within my datasets I have both classification and regression models which I must train a multi-layer perceptron. For these I have used scikit-learn’s MLPClassifier[64] and MLPRegressor[65] they both use backpropagation to train a multi-layer perceptron. The regression model differs by not having an activation function in the final output layer which allows its output to be continuous [63]. They also offer a both a RandomForestClassifier[] and RandomForestRegressor[] for the random forest components of testing.

Neural networks require their inputs to be numerical, this presents an issue for categorical data. one could simply represent each category as a number however this infers an order between categories where there might not be one. This can lead to poor performance, and so unless your categories are in fact ordinal data then this should be avoided, another method is a one hot encoding, which makes an attribute for each category and represents it as a Boolean. This Is a common technique which is effective for small number however increases the sparsity of the data set. Having features with a varying range can result in features with a high-range having outsized impact on the prediction for this reason I will normalise numerical input with min-max scaling. This will ensure that all numbers lie between 0 and 1 keeping their range constant.

Both of these data preparation steps can be applied using sklearn.preprocessing package. Which can be added as steps in the sklearn.pipeline, this allows the preparation steps to be saved alongside the model. this becomes extremely useful when we reuse our model as it comes with all the preparation steps required to pass in the raw data. especially for our purposes of passing the model to Dice and generating data to feed into it.

## Measurement

### Classification

I will use AUC ROC [40]. I have selected this over accuracy alone as it gives a broader description of a model [40] and because we have dealt with data imbalance, we don’t need the sensitivity provided by an F1 score [41]. For multivariant classification I will apply one against all method [39] and take the mean of the score.

### Regression

For regression I will use mean squared error (MSE). Correlation coefficient

### Complexity

Interpretability of a tree is negatively correlated with its depth [43]. Ribeiro [6] calculates the complexity of a decision tree as count of its leaf nodes, whilst Saadallah [8] determines the average decision path length, they reason that the number of leaf nodes method penalizes shorter decision paths. I use average path length to compute complexity of my trees.

## Experiment 1 – Global Fidelity

***Hypothesis0:*** *A global surrogate model trained on CF examples will have no greater accuracy relative to a global surrogate model trained on randomly selected values.*

To test the method without the influence of locality I will take a random sample from the global feature space of N samples. I will compare the relative accuracy between the CF method and the randomised control which will consist of the same randomised sample used as the input for the CF method [fig 2]. I will repeat the experiment, raising the number of samples as a dependent variable and track accuracy, time taken and complexity as dependant. If I observe a statistically significant increase in accuracy compared to the control method I will be able to reject my null hypothesis.

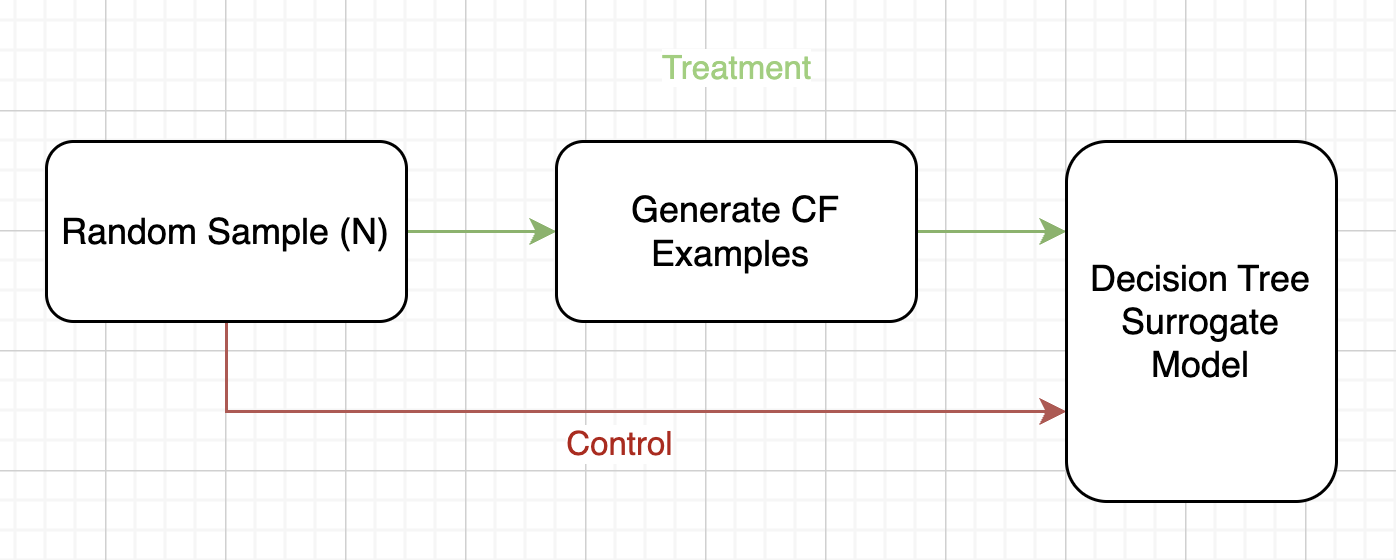


Figure : experiment 1 control and treatment diagram

## Experiment 2 – Local fidelity

***Hypothesis0:*** *A Local surrogate model trained on CF examples will have no greater accuracy relative to a global surrogate model trained on randomly selected values.*

Diagram

Description automatically generated

Figure : Experiment 2 flow diagram showing treatment and control

I will use the Pearson correlation to show if there is a correlation between N and the accuracy of the model. I expect from [13] to see a positive correlation in my randomised control group however, I hypothesis that my method will have a stronger correlation than the control.

## Experiment 3 – Counterfactual

To determine how

# Results

## Hardware

All tests were performed on the same hardware with a 32GBs of memory and a AMD 8 Core - 3800 MHz processor. However due to python’s Global Interpreter Lock [69] the processor is not able to fully utilize its multicore advantage. This does not mean the code could not benefit from improvement to make better use of threads however this is beyond the scope of this project.

## Models Under Analysis

As previously mentioned in the method section, the accuracy scores for the model under analysis

|  |  |  |  |
| --- | --- | --- | --- |
| Model Name | Problem Type | Metric | Score |
| Iris-nn | Classification | AUC Score | 1.0 |
|  |  | Accuracy | 1.0 |
| dry-bean-nn | Classification | AUC Score | 0.9944 |
|  |  | Accuracy | 0.9243 |
| Adult-nn | Classification | AUC Score | 0.8666 |
|  |  | Accuracy | 0.8249 |
| bike-sharing-hourly-nn | Regression | Mean squared error | 1707.13 |
|  |  | Coefficient of determination  (R-squared) | 0.95 |
|  |  | Mean Absolute Error | 25.80 |
| wine-quality-nn | Regression | Mean squared error | 0.47 |
|  |  | Coefficient of determination  (R-squared) | 0.36 |
|  |  | Mean Absolute Error | 0.53 |

## Experiment 1 - Global Surrogate Modal

### Time analysis

The first thing that became obvious whilst running the experiment was the time to run the experiment limited N samples dependant variable as the optimization process for generating counterfactuals proved to be a resource intensive task especially for task multiple classification were the

<graph>

### Performance

<Accuraccy VS N\_samples>

Demsar [38] warns against using the parametric t-test for the analysis of machine learning across multiple datasets due to them being incommensurable and their assumption of a normal distribution. They instead suggest Wilcoxon signed-ranks test which is a non-parametric version of the T-test. This will help us detect if there is a statistically significant difference ( α = 0.5 ) between our baseline and treatment observations.

# Bibliography

[1]  Rudin C, Wang C, Coker B. The age of secrecy and unfairness in recidivism prediction. arXiv e-prints 1811 00731 [applied statistics]. 2018 Nov;.

[2] Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan, “Dissecting racial bias in an algorithm used to manage the health of populations,” *Science*, vol. 366, no. 6464, pp. 447–453, 2019.

[5] J. Larson, J. Angwin, L. Kirchner, and S. Mattu, “How we analyzed the compas recidivism algorithm,” *ProPublica*, 23-May-2016. [Online]. Available: https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm. [Accessed: 07-Jan-2023].

[6] M. T. Ribeiro, S. Singh, and C. Guestrin, “‘why should I trust you?",” *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016.

[7] S. PIRAMUTHU, “Input data for decision trees,” *Expert Systems with Applications*, vol. 34, no. 2, pp. 1220–1226, 2008.

[8] A. Saadallah and K. Morik, “Active sampling for learning interpretable surrogate machine learning models,” *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*, 2020.

[9] M. Phuong and C. H. Lampert, “Towards understanding knowledge distillation,” *arXiv.org*, 27-May-2021. [Online]. Available: https://arxiv.org/abs/2105.13093. [Accessed: 07-Jan-2023].

[10] J. Song, H. Zhang, X. Wang, M. Xue, Y. Chen, L. Sun, D. Tao, and M. Song, “Tree-like decision distillation,” *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

**[11]** C. Molnar, *Interpretable machine learning: A guide for making Black Box models explainable*. Munich, Germany: Christoph Molnar, 2022.

**[12]** C. Rudin, “Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead,” *Nature Machine Intelligence*, vol. 1, no. 5, pp. 206–215, 2019.

[13]  O. Bastani, C. Kim, and H. Bastani, “Interpretability via model extraction,” *arXiv preprint arXiv:1706.09773*, 2017.

[14] P. Chandrasekar, K. Qian, H. Shahriar and P. Bhattacharya, "Improving the Prediction Accuracy of Decision Tree Mining with Data Preprocessing," 2017 IEEE 41st Annual Computer Software and Applications Conference (COMPSAC), 2017, pp. 481-484, doi: 10.1109/COMPSAC.2017.146.

[15] S. B. Kotsiantis, “Decision trees: A recent overview,” *Artificial Intelligence Review*, vol. 39, no. 4, pp. 261–283, 2011.

[16] R. Kommiya Mothilal, D. Mahajan, C. Tan, and A. Sharma, “Towards unifying feature attribution and counterfactual explanations: Different means to the same end,” *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 2021.

[17] Brent Mittelstadt, Chris Russell, and Sandra Wachter. 2019. Explaining Explanations in AI. In Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT\* '19). Association for Computing Machinery, New York, NY, USA, 279–288. https://doi.org/10.1145/3287560.3287574

[18] S. Wachter, B. Mittelstadt, and C. Russell, “Counterfactual explanations without opening the black box: Automated decisions and the GDPR,” *SSRN Electronic Journal*, 2017.

[19] Ramaravind K. Mothilal, Amit Sharma, and Chenhao Tan (2020). **Explaining machine learning classifiers through diverse counterfactual explanations**. Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency.

[20] Model-Agnostic Counterfactual Explanations for Consequential Decisions A.-H. Karimi, G. Barthe, B. Belle, I. Valera, Model-agnostic counterfactual explanations for consequential decisions, arXiv preprint arXiv:1905.11190 (2019).

[21] Interpretml, “Interpretml/dice: Generate diverse counterfactual explanations for any machine learning model.,” *GitHub*. [Online]. Available: https://github.com/interpretml/DiCE. [Accessed: 07-Jan-2023].

[22] T. Freiesleben, “The intriguing relation between counterfactual explanations and adversarial examples,” *Minds and Machines*, vol. 32, no. 1, pp. 77–109, 2021.

[23] R.-H. Li and G. G. Belford, “Instability of decision tree classification algorithms,” Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '02, 2002.

[24] H. Blockeel and J. Vanschoren, “Experiment databases: Towards an improved experimental methodology in machine learning,” *Knowledge Discovery in Databases: PKDD 2007*, pp. 6–17.

[25] J. Zobel, *Writing for computer science: The art of effective communication*. Singapore: Springer, 2000.

[26] C. Rudin, C. Chen, Z. Chen, H. Huang, L. Semenova, and C. Zhong, “Interpretable machine learning: Fundamental principles and 10 grand challenges,” *Statistics Surveys*, vol. 16, no. none, 2022.

[27] Kaufman, S., Rosset, S., Perlich, C., and Stitelman, O. 2012. Leakage in data mining: Formulation, detection, and avoidance. ACM Trans. Knowl. Discov. Data. 6, 4, Article 15 (December 2012), 21 pages.

[28] J. Reichertz, “Induction, deduction, abduction,” The SAGE Handbook of Qualitative Data Analysis, pp. 123–135.

[29] J. Creswell W. and J. Creswell David, Research Design. SAGE Publications, 2017.

[30] Stodden V, Guo P, Ma Z (2013) Toward Reproducible Computational Research: An Empirical Analysis of Data and Code Policy Adoption by Journals. PLoS ONE 8(6): e67111. doi:10.1371/journal.pone.0067111

[31] UCI Machine Learning Repository: Iris data set. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Iris/. [Accessed: 07-Jan-2023].

[32] UCI Machine Learning Repository: Adult. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Adult/. [Accessed: 07-Jan-2023].

[33] UCI Machine Learning Repository: Dry Bean Dataset. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/ Dry+Bean+Dataset/. [Accessed: 07-Jan-2023].

[34] UCI Machine Learning Repository: Wine Quality. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Wine+Quality/. [Accessed: 07-Jan-2023].

[35] UCI Machine Learning Repository: bike sharing dataset. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/bike+sharing+dataset/. [Accessed: 07-Jan-2023].

[36] Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

[37] D. A. Cieslak and N. V. Chawla, “Learning decision trees for unbalanced data,” *Machine Learning and Knowledge Discovery in Databases*, pp. 241–256.

[38] Demšar, Janez. "Statistical comparisons of classifiers over multiple data sets." *The Journal of Machine learning research* 7 (2006): 1-30.

[39] Allwein, E.L., Schapire, R.E. and Singer, Y. (2000). Reducing multiclass to binary: A unifying approach for margin classifiers. *Journal of Machine Learning Research*, **1**:113–141.

[40] Ling, Charles X., Jin Huang, and Harry Zhang. "AUC: a statistically consistent and more discriminating measure than accuracy." *Ijcai*. Vol. 3. 2003.

[41] Sokolova, Marina, Nathalie Japkowicz, and Stan Szpakowicz. "Beyond accuracy, F-score and ROC: a family of discriminant measures for performance evaluation." *Australasian joint conference on artificial intelligence*. Springer, Berlin, Heidelberg, 2006.

[43] Izza, Yacine, Alexey Ignatiev, and Joao Marques-Silva. "On explaining decision trees." *arXiv preprint arXiv:2010.11034* (2020).

[44] “Sklearn.ensemble.randomforestclassifier,” scikit. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html. [Accessed: 07-Jan-2023].

[45] “Decision trees,” *scikit*. [Online]. Available: https://scikit-learn.org/stable/modules/tree.html. [Accessed: 07-Jan-2023].

[46] “Neural network models (supervised),” *scikit*. [Online]. Available: https://scikit-learn.org/stable/modules/neural\_networks\_supervised.html. [Accessed: 07-Jan-2023].

[47] “Creative Commons License Deed,” *Creative Commons - Attribution 4.0 International - CC BY 4.0*. [Online]. Available: https://creativecommons.org/licenses/by/4.0/. [Accessed: 07-Jan-2023].

[48] “Code of practice on ethics,” *Code of practice on ethics - Staff home, University of York*. [Online]. Available: https://www.york.ac.uk/staff/research/governance/research-policies/ethics-code/. [Accessed: 07-Jan-2023].

[49] D. Wang et al., "Nonlinear decision boundary created by a machine learning-based classifier to mitigate nonlinear phase noise," 2015 European Conference on Optical Communication (ECOC), 2015, pp. 1-3, doi: 10.1109/ECOC.2015.7341753.

[50] Bianca Zadrozny. 2004. Learning and evaluating classifiers under sample selection bias. In Proceedings of the twenty-first international conference on Machine learning (ICML '04). Association for Computing Machinery, New York, NY, USA, 114. <https://doi.org/10.1145/1015330.1015425>

[51] G. D. Ruxton and M. Neuhäuser, “When should we use one-tailed hypothesis testing?,” Met

**[52]** M. T. Ribeiro, S. Singh, and C. Guestrin, “‘why should I trust you?",” *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016.

[53] O. I. Abiodun, A. Jantan, A. E. Omolara, K. V. Dada, N. A. E. Mohamed, and H. Arshad, “State-of-the-art in Artificial Neural Network Applications: A survey,” Heliyon, vol. 4, no. 11, 2018.

[54] K. Hornik, M. Stinchcombe, and H. White, “Multilayer feedforward networks are universal approximators,” *Neural Networks*, vol. 2, no. 5, pp. 359–366, 1989.

[55] Chris M. Bishop , "Neural networks and their applications", Review of Scientific Instruments 65, 1803-1832 (1994) <https://doi.org/10.1063/1.1144830>

[56] G. Sanderson, “Gradient descent, how neural networks learn | Chapter 2, Deep Learning,” *YouTube - 3Blue1Brown*, 16-Oct-2017. [Online]. Available: https://www.youtube.com/watch?v=IHZwWFHWa-w. [Accessed: 27-Feb-2023].

[57] Breiman, L. Random Forests. Machine Learning 45, 5–32 (2001). https://doi.org/10.1023/A:1010933404324

[60] Sophia Daskalaki, Ioannis Kopanas & Nikolaos Avouris (2006) EVALUATION OF CLASSIFIERS FOR AN UNEVEN CLASS DISTRIBUTION PROBLEM, Applied Artificial Intelligence, 20:5, 381-417, DOI: 10.1080/08839510500313653

[61] SMOTE: Synthetic Minority Over-sampling Technique

[63] https://scikit-learn.org/stable/modules/neural\_networks\_supervised.html

[64] https://scikit-learn.org/stable/modules/generated/sklearn.neural\_network.MLPClassifier.html

[65] https://scikit-learn.org/stable/modules/generated/sklearn.neural\_network.MLPRegressor.html

[66] Defining Locality for Surrogates in Post-hoc Interpretablity, <https://arxiv.org/pdf/1806.07498.pdf>

[67] <https://numpy.org/doc/stable/reference/random/generated/numpy.random.normal.html>

[68] <https://numpy.org/doc/stable/reference/random/generated/numpy.random.choice.html>

[69] https://wiki.python.org/moin/GlobalInterpreterLock