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 2](#_Toc129790753)

[1.1 Focus of Research 2](#_Toc129790754)

[1.2 Question 1 2](#_Toc129790755)

[1.3 Question 2 2](#_Toc129790756)

[1.4 Question 3 2](#_Toc129790757)

[2 Literature Review 3](#_Toc129790758)

[1.5 Neural Networks 3](#_Toc129790759)

[1.6 Random Forests 3](#_Toc129790760)

[1.7 Explainable AI 4](#_Toc129790761)

[1.7.1 Surrogate Models 4](#_Toc129790762)

[1.7.2 LIME 4](#_Toc129790763)

[1.7.3 Decision trees 5](#_Toc129790764)

[1.7.4 Counterfactuals 6](#_Toc129790765)

[3 Theory 8](#_Toc129790766)

[4 Research Methodology 9](#_Toc129790767)

[1.8 Epistemology 9](#_Toc129790768)

[1.9 Experiment Design 10](#_Toc129790769)

[1.10 Preparing black box models 10](#_Toc129790770)

[1.10.1 SMOTE 10](#_Toc129790771)

[1.10.2 Sklearn 10](#_Toc129790772)

[1.11 Generating counterfactuals 11](#_Toc129790773)

[1.11.1 Multi-class classification 11](#_Toc129790774)

[1.11.2 Regression 11](#_Toc129790775)

[1.12 Locality of explanations 13](#_Toc129790776)

[1.13 Experiment 1 13](#_Toc129790777)

[1.14 Experiment 2 14](#_Toc129790778)

# 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 aim to develop a model agnostic post-hoc technique which can be used to improve the fidelity of a surrogate model whilst also providing clarity.

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 explore Wachter’s [18] counterfactual optimization applied to improving the accuracy of a surrogate model.

# Questions

### Question 1

Can training a surrogate model on counterfactual examples improve the fidelity/accuracy?

**H0 :** There is no increase in accuracy in the surrogate models trained on counterfactual examples compared to the randomised sampling.

**H1 :** There is an increase in accuracy in the surrogate models trained on counterfactual examples compared to the randomised sampling.

### Question 2

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

### Question 3

How can the surrogate models offer insights into the structure of the black box?

# Literature Review

## Blackboxes

## 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. (maybe discuss). 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]

By default, Lime achieves Locality through the radial basis function which plots distance in hyperdimensional space. It generates hypothetical instances within a given kernel length of the a given example. Molnar [11] warns about 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. Laugel et al [66] discus different methods for establishing an area to localise perturbations, they compare limes kernel 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.

CART Decision trees use a greedy algorithm to split data by rules that minimize the Gini impurity [15]. 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].

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 [Equation1]

Equation. 1

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 watchter [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 minimize their proximity to the decision boundary, I posit that similar to 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)

The *Rashmond effect is of* benefit as multiple CF examples with equal distance will be identified, this can be controlled through the diversity and proximity parameters.

A picture containing shape

Description automatically generated

Figure 1 : 2 generations of CF examples and their relation to the decision boundary

# Research Methodology

## Epistemology

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

I seek to understand an ontologically deterministic process, I believe 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).

I have formalised Question 1 as a hypothesis which can be used to make observations. This hypothesis is one-tailed [51], in that it describes only two potential outcomes. Question 2 requires a similar approach however, until I have conducted further investigation, I will not be able to formulate a robust hypothesis.

I intend to test this method quantitively through experimental observation, 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.

The third question has a qualitative component, it seeks to evaluate the usefulness of the surrogate models to gain insight, which is subjective. I will take a mixed methods explorative approach [29] by both quantifying their complexity and exploring these trees through my own emic perspective as practitioner trying to discern patterns from the backbox models which I can display as examples. This approach is shared by [8],[13],[6].

## Experiment Design

## 

Diagram

Description automatically generated

Figure 2: Overview of experiment stages

## Preparing black box models

Blockeel advocate for “varying the conditions under which [machine learning research] experiments are run” [24] to ensure results are generalizable. I run the experiment on 5 different open datasets which are representative of common machine learning tasks and construct both neural network and a random forest model for each.

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 the models.

### SMOTE

The Adult Dataset has a large amount of class imbalance, it has been shown that class imbalance can impact prediction by favouring the majority class [60].

One approach would be to under sample the majority class however this comes at the cost of limiting the number of training examples. Synthetic Minority Oversampling Technique (SMOTE) [61] is a technique that uses k nearest neighbour to generate training examples for the minority class that are close in the feature space to the original data [61].

### Sklearn

For this research’s predictive 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 randomforest 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 which 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.

It is common practice to pick the number of layers and weights through trial and error by adjusting the network until the loss improves.

(whilst this isn’t the most efficient search strategy it suits our purposes.)

## Generating counterfactuals

Traditional counterfactuals explanations are designed to be an explanation of a single value however in this research I will generate a large volume of them for a verity of problem types.

### Multi-class 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 this however this results in slightly less than n^2 comparisons. To constrain this, I randomly sample a percentage relative to the number of classes to keep n constant. I ensure the sample is stratified to ensure that an equal number of each class is represented.

Text

Description automatically generated

Figure : Multiclass counterfactual generation (I have stripped away unnecessary implementation details)

### 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.

{for lit review}

Lime uses a kernel function to generate

A radial kernel allows us to rank similarity of two instances in infinite dimensional space, thus we can generate random perturbations with that have a with a low kernel width

Lime generates training data for its surrogate model through perturbating on examples which It generates through.

Localisation

Lime generates perturbations to train on by taking a sample that follows the same characteristics of the feature space (standard deviation and mean)

[https://arxiv.org/pdf/1806.07498.pdf], it ranks values through a kernel function and minimising distance to the explanatory value.

The use of clustering for localisation.

[] show that this

[] explored the effect that neighbourhood selection has on fidelity, they found that LIMEs [Ribeiro] default sampling technique suffered from global features masking the effect of locally important features. They suggested the use of generate.

Lime The samples have the same mean and standard deviation as the original feature space X

## Locality of explanations

Why do we do it ? (maybe in the lit-view)

Localisation is achieved through the nearest neighbour algorithm, which utilizes k means clustering to represent all features in a hyperdimensional space and computes **Minkowski distance between a given example and the other instance. Minkowski is a generalised form of Euclidean distance (p=2).** One method for selecting nearest neighbours 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 most 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 values 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 a large amount of categorical values such as the adult dataset.

After establishing the subset, the counterfactuals of each are generated for two generations.

Read { https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6822444 }

KNN doesn’t perform well in high dimentional data (peeking phenomenom)

[https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.418.6517&rep=rep1&type=pdf]

## Experiment 1

To test the hypothesis of Question 1, I will then train a

Cart Decision tree on each of the datasets, controlling N between dataset1 anddataset2.. If I observe a statistically significant increase in accuracy compared to the baseline, I will be able to reject my null hypothesis. 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.

## Experiment 2

To investigate the effect of proximity and diversity on accuracy I will have to test my assumption about their effect in respect to the number of training instances generated as this unknown and will depend on the nature of the Blackbox. I will do this by incrementally increasing proximity and diversity in the counterfactual generation stage. Once I have established that there is substantially more training Instance due to the Rashmond effect [11]. I will repeat the experiment 1 gradually increasing N training instances.

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.

# Results

Global Surrogate Modal

After applying the method globally on the entire training set I found the method takes far too long to generate counterfactuals to be useful as a diagnostic tool. As can be seen []

Accuracy :

<table of resutls>

Resulting Tree

Local Surrogate Model

[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