Explaining Credit Card Fraud Decisions in ML: An Analysis of XAI Methods

Ciaran Finnegan - D21124026

A dissertation submitted in partial fulfilment of the requirements of
Dublin Institute of Technology for the degree of
M.Sc. in Computing (Data Science)

March 2024

Declaration

I certify that this dissertation which I now submit for examination for the award of MSc in Computing (Data Science), is entirely my own work and has not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the text of my work.

This dissertation was prepared according to the regulations for postgraduate study of the Technological University Dublin and has not been submitted in whole or part for an award in any other Institute or University.

The work reported on in this dissertation conforms to the principles and requirements of the Institute's guidelines for ethics in research.

Signed:

Date:

Abstract

The Covid pandemic accelerated an already rapidly evolving trend towards a cashless society, and simultaneously led to an even greater demand for more effective online fraud protection measures. The Financial Services industry needs to maintain its own momentum in fraud prevention by screening online transactions with increasingly sophisticated Machine Learning algorithms. At the same time, these service providers do not have a free hand in the implementation of fraud prevention applications, as regulators and public alike still demand accountability from industry AI. This is a theme which looms even larger as the presence of GenAI becomes even more ubiquitous in the modern world, and a certain suspicion around the perceived reliance in AI grows as the first quarter of the 21st century ends.

Credit card fraud detection through ML technology has been a commonplace and active area of research for nearly two decades. This is heavily driven by the persistent and mounting threat from *'bad actors'* in this domain, and the billions of Euros lost each year by individuals and Financial Institutions. Access to better sources of data and more sophisticated models continues to improve card fraud detection rates. However, the challenges from criminals continues to evolve, and financial institutions increasingly rely on concepts such as Artificial Neural Networks (ANN) to increase the speed and accuracy of credit card fraud detection. Even a cursory review of academic research in the area of credit card fraud prevention will shine a light on the dilemma that companies face in this particular domain of crime prevention. It is not enough to stop a case of suspected credit card fraud; a company must be able to explain *why*. The workings of ANN models are not readily apparent and trust in the *'black box'* alone will not suffice. Also, increasingly, it is not just a case of justifying a decision on

fraud to an individual customer; as industry and government regulators will likewise demand that such decision-making processes are transparent and comprehensible.

Companies delivering applications to the financial crime prevention business sector have been cognisant in very recent years of the need to add explainability into their product suite. Vendors (such as IBM, Actimize, SymphonyAI, etc.) will boast about advancements in detection rates in areas such as credit card fraud, but have also started to supplement these offerings with built-in explanation data for fraud investigators. Driven by regulatory requirements, company auditors will demand that Financial Institutions can demonstrate an developing process to prevent credit card crime but can also stand over decisions as to why a client's cards transaction has been delayed/rejected (or why not). That said, commercial Product development of embedded explanations in fraud detection tools is still at a relatively nascent stage.

Data Scientists working in this commercial sphere are aware of the various Explainable Artificial Intelligence (XAI) techniques that can be applied in the domain of credit card fraud, but there is still relatively little published research on the comparative benefits of such approaches in this context. Furthermore, a significant volume of research focuses on the human interpretation of XAI output, whether the subject is fraud detection or health care prediction. Human surveys are costly to implement and can be susceptible to the bias and/or lack of domain knowledge of the participant. The focus of this paper is to look at an automated and statistical comparison of established XAI methods for credit card fraud and assess if a quantitative difference exists between them in terms of general performance. Such an analysis could provide guidance to future product roadmaps in the commercial online fraud prevention space.

Keywords: Explainable Artificial Intelligence, Credit Card Fraud Detection, Interpretability, XAI Statistical Comparison

Acknowledgments

I would like to express my sincere thanks to Dr Bujar Raufi and Dr Luco Longo for their academic guidance with this dissertation. In addition I wish to acknowledge my colleagues in SymphonyAI; Dr Rory Duthie, Eddie Baggot, Kieran MacKenna and Dan Branley for their suggestions and provocations.

Dataset: Sourced, and used with permission, from 2015 product research conducted by Norkom Technologies on emerging fraud detection techniques.

Contents

List of Figures

List of Tables

List of Acronyms

XAI	Explainable Artificial Intelligence
ANN	Artificial Neural Network
SHAP	SHapley Additive exPlanations
LIME	Local Interpretable Model-agnostic Explanations...
DiCE	Diverse Counterfactual Explanations

Chapter 1

Introduction

1.1 Background

Credit card fraud costs the Financial Services industry billions of Euros in losses each year (Nesvijevskaia, Ouillade, Guilmin, & Zucker, 2021). The need for ever more sophisticated Machine Learning techniques to tackle this problem has been well established by academic observers such as (Dal Pozzolo et al., 2014) and (P. Sharma & Priyanka, 2020). Research by (A. Sharma & Bathla, 2020) and (Batageri & Kumar, 2021) are examples of work in this field to improve fraud detection rates through ever more sophisticated neural network algorithms. However, many researchers highlight the parallel challenge that these *'black box'* models need to be held accountable for the individual fraud classifications that they make (T.Y.Wu & Y.T.Wang, 2021).

(Ignatiev, 2020) focuses on the need for Explainable Artificial Intelligence (XAI) to be *trustable*, while (Carvalho, Pereira, & Cardoso, 2019) are more emphatic about the European Union's legal demands that all automated decision making about citizens be *transparent*.

This dissertation will focus on ML driven software used by the Financial Services industry and whether an objective rating can be given to different XAI methods in terms of explaining the reason for a given credit card fraud classification. To narrow the field of interest further, the paper will propose a series of metrics to rate the performance of four state-of-the-art XAI methods; SHAP, LIME, ANCHORS, and DICE

on an industry credit card fraud dataset, as applied to the classification of individual credit card transactions. (These XAI approaches are described, with supporting references, in Section 2.2). Companies operating in the area of financial crime software, such as SymphonyAI and Actimize, already sell ML based software to detect credit card fraud but generally rely on only one explainer technique, usually SHAP values.

Specifically, the scope of experiments in this paper is on explanations for individual (*'local'*) transactions, and only considers interpretability techniques that are *agnostic* about the type of the detection model.

1.2 Research Project/Problem

1.2.1 Research Question

"To what extent can we quantify the quality of contemporary machine learning interpretability techniques, providing local, model-agnostic, and post-hoc explanations, in the classification of credit card fraud transactions by a 'black box' Neural Network ML model?"

The question focuses on a quantitative comparison of explanations produced by different XAI techniques on specific (local) NN model predictions.

1.2.2 Research Problem

The research problem can be described as the means to produce an objective assessment of state-of-the-art ML explainers, as applied to credit card fraud detection. The intention is to compare a set of common XAI techniques and look for insights into the relative strengths of each one. The initial experiment focus is on the application of SHAP, LIME, ANCHORS, and DiCE interpretability methods upon a Neural Network model trained on a commercial dataset containing credit card transactions, which are labelled *'fraud'* or *'non-fraud'*.

1.3 Research Objectives

Metrics for all four explainer techniques will be collated based on the methodology described in Section 1.4 below. This output will be subjected to a statistical test for significance. Is one explainer better than another and if so, how great is that difference?

The objective of the experiments under-pining this research paper is a very deliberate intention to avoid any elements of human assessment of the XAI techniques. Taking a lead from a study of XAI techniques applied to eye tracking experiments in the research by (Martínez, Nadj, Langner, Toreini, & Maedche, 2023) this research will aim to demonstrate that it is possible to produce a purely statistical analysis on explainer performance for individual credit card fraud predictions.

1.4 Research Methodologies

The proposed experiments in this paper are based on a similar study into measuring interpretability methods on healthcare datasets that classified mortality predictions (ElShawi, Sherif, Al-Mallah, & Sakr, 2020). This study defined a set of generic metrics and then *scored* the output of a series of XAI techniques against these metrics.

As listed in Section 1.1, the experiments in this research paper will score SHAP, LIME, ANCHORS, and DICE methods on an industry credit card fraud dataset. The metrics for the experiments are defined in Section 3.2.2 of this paper and take inspiration from the aforementioned (ElShawi et al., 2020) research and XAI benchmarks produced in a paper by (Jacob et al., 2021).

A key assumption is that this research approach will translate into the domain of credit card fraud detection.

1.5 Scope and Limitations

Focus on Four Specific XAI Techniques

Experiments are being specifically limited to four post-hoc and local interpretability frameworks. Thus the research is only looking at explanations extracted from a trained model for individuals credit card fraud predictions. This is done to focus on the 'business' objectives of this paper, as elaborated in the Abstract, and in particular to build on related research papers by (Ribeiro, Singh, & Guestrin, 2016) and (Guidotti et al., 2019).

Only Local Explanations Measured

Only local explanations on specific credit card transactions are being considered – global explainability on the overall model is not in scope. The potential use case for the output of this research is to improve fraud investigator information for specific transaction assessments. The choice of explainers for this paper has also been influenced by the graphical display of the most common XAI techniques for local assessment of NN models provided in research by (Ras, Xie, Gerven, & Doran, 2022).

Automated Experiments

Deliberately, there is no human assessment of the explanations as this will be a purely programmatic and arithmetic exercise. This is a conscious research decision to implement a purely statistical analysis. The rationale is to generate a lower cost assessment framework for XAI output but to also avoid situations such as those described by (Chromik, .Eiband, Buchner, Krüger, & Butz, 2021) whereby users surveyed fall foul of that the authors describe as *"...the illusion of Explanatory Depth in Explainable AI..."*.

Potential Hardware Limitations

If the use of extensive GPU processing is required for certain explainers, then this may be beyond what can be afforded this dissertation, and experiment scope may have to

be reduced.

1.6 Document Outline

This dissertation begins with a critical review of relevant recent research conducted in the area of Explainable AI (XAI), with particular focus on the evaluation of common explainer methods as applied to model predictions. The research topic of this paper is strongly focused around techniques for local model-agnostic interpretations of classification model results, and that has informed the literature reviewed and assessed in the following chapter (Chapter 2).

Chapter Three describes the experimental approach to build an evaluation matrix for the four chosen explainer techniques; SHAP, LIME, ANCHORS, and DiCE (Counterfactuals). Five sets of metrics are generated based on the explanation outputs generated by each technique on predictions for a binary credit card fraud classification problem.

The methodologies described in Chapter Three deliver the XAI metrics outputs, which are then assessed in Chapter Four. A significance test is conducted to evaluate the relative effectiveness of each technique. This analysis is supplemented by a further statistical calculation as to whether these approaches can be ranked in any order of merit.

The paper concludes in Chapter 5 with a consolidation of these results and the possible future application of the analysis, along with other recommended avenues for investigation.

Chapter 2

Review of existing literature

2.1 Key Themes in Current Research

2.1.1 How to Measure the Effectiveness of an Explanation? No Obvious Consensus

The literature review for this dissertation began with assessments of how the detection of credit card fraud by Machine Learning models is being refined with ever more sophisticated neural network models (P. Sharma & Priyanka, 2020). However, in their research experiments with the LIME algorithm, (Ribeiro et al., 2016) describe how users can have a trust issue with such ML models, like NN, because they are effectively *'black-boxes'* from which it is very difficult to interpret why a given classification has been derived. This is a theme echoed in the introduction to many research papers, such as (ElShawi et al., 2020), (Honegger, 2018), and (Sinanc, Demirezen, & Sağıroğlu, 2021). Despite these acknowledgement, in this research domain there appears to be no cast iron process to establish XAI *trustworthiness*. Although attempts at building universal frameworks to interpret model predictions have been proposed (Lundberg & Lee, 2017) there is still no unanimity seen in research to date on what constitutes an objectively *'good'* explanation of a prediction. The gap remains; how exactly does a researcher measure and display *'explainability'* in Explainable Artificial Intelligence (XAI) research?

To add further emphasis on this gap in contemporary research, (Adadi & Berrada, 2018) claimed that *"Technically, there is no standard and generally accepted definition of explainable AI"* (p. 141). More specifically, in their review of XAI research papers, (Vilone & Longo, 2021b) state that *"There is not a consensus among scholars on what an explanation exactly is, and which are the salient properties that must be considered to make it understandable for every end-user."* (p.651) Therefore, as stated above, there is no well-established output framework for explaining credit card fraud classification through *'black-box'* models (Vilone & Longo, 2021a).

This paper proposes to build on some of the objective research on scoring predictions generated by four established interpretability methods.

2.1.2 Human Assessment vs Automated Benchmarks

Research by (Jacob et al., 2021) makes the point about assessing XAI output that *"...while a user study may be the best way to evaluate the usefulness of explanations, it is not always available and may come at a high cost."*. It is also desirable for humans taking part in XAI surveys to have some degree of domain knowledge, but fraud detection explainer experiments by (Jesus et al., 2021) showed that this can still be subject to user bias.

Examples of XAI research where the reliance on human assessment of explanations is less commonplace can be seem in the domain of healthcare, through research by (Marcilio & Eler, 2020) and (Lakkaraju, Bach, & Leskovec, 2016). Those experiments produce clearly objective recommendations in line with the work of (ElShawi et al., 2020). Research into explanations for ML fraud classification often follow a more subjective survey style of experimentation involving the augmentation of human based processes with model explainer outputs. On occasion, human bias can adversely impact on the reliability of the interpretation of the ML generated model explanations (Kaur et al., 2020). This dissertation will follow in the steps of earlier research that only use non-human, programmatic experiments with quantifiable metrics (Darias, Caro-Martínez, Díaz-Agudo, & Recio-Garcia, 2022) and tests for statistical significance (Evans, Xue, & Zhang, 2019).

The methodology for this paper's experiments are based heavily on the 2020 healthcare XAI research by Elshawi et al, but also take inspiration from the anomaly detection framework for explainability created by (Nguyen et al., 2023) and XAI time series results produced by (Schlegel, Arnout, El-Assady, Oelke, & Keim, 2019), both using SHAP and LIME explainer outputs. A further bespoke framework can be seen in research by (Moreira et al., 2020) using local explanations also generated with SHAP, LIME and Counterfactual techniques.

The use of statistical significant tests, which form the basis of the experiment assessments in Section 4.2 of this paper, follow the ML evaluation processes described by (Hanafy & Ming, 2022).

The metrics described in Section 3.2.2 of this paper are also influenced by the aforementioned healthcare research but have been augmented by concepts of 'robustness' expounded on by (Alvarez-Melis & Jaakkola, 2018)

2.1.3 Neural Networks and XAI

In order to reflect that credit card fraud detection models are relying on increasingly more sophisticated NN algorithms (Ajitha, Sneha, Makesh, & Jaspin, 2023) (Aurna, Hossain, Taenaka, & Kadobayashi, 2023), the experiments in this paper will involve local *post-hoc* explanations generated on a trained NN model. A significant body of research material on XAI approaches using Deep Learning (DL) techniques was assessed for this paper. This helped direct the experiments described in Section 3.2 on the use of procedures such as Deep SHAP (Nascita et al., 2021) (Sullivan & Longo, 2023), LIME for Deep Neural Network (DNN) experiments (Ras et al., 2022), and the use of counterfactuals in CNN model explainers (Vouros, 2022).

2.1.4 Computational Efficiency

Also of note is the observation from (Psychoula et al., 2021) that the runtime implications of XAI output on real-time systems, fraud or otherwise, has had relatively little research focus to date. Early prototyping in this dissertation effort will attempt to

capture and address any such issues as quickly as possible.

2.1.5 Presenting XAI Data

(Guidotti et al., 2019) conducted comparative experiments into local interpretability frameworks but note in their conclusions that is still relatively little research into building more aesthetically attractive visualisations of such explanations. This will not be a focus area of this dissertation.

2.2 State of the Art Approaches for Local Interpretability

This section of the document describes research conducted into the local interpretability techniques that formed the basis of the experiments in this dissertation research.

2.2.1 SHAP

SHAP stands for **SH**apley **A**dditive ex**P**lanations (Lundberg & Lee, 2017) and can be described as a unified framework for interpreting predictions. It provides a toolkit that is computationally efficient at calculating 'Shapley' values. SHAP is a method derived from cooperative game theory and SHAP Values are used extensively to present an understanding of how the features in a dataset are related to the model prediction output. It is a *'black box'* explainability technique that can be applied to most algorithms without being aware of the exact model.

The focus of this dissertation research is on local interpretations, so we will be using SHAP to understand how the NN model made a fraud classification for a single transaction instance. (SHAP values can also be used for global interpretations of a given model).

2.2.2 LIME

LIME stands for **L**ocal **I**nterpretable **M**odel-agnostic **E**xplanations (Ribeiro et al., 2016) and is also a popular choice for interpreting the decisions made by black box models. The core concept of LIME is that it aims to understand the features that influence the prediction of a given black box model around a single instance of interest. LIME approximates these predictions by training local surrogate models to explain individual predictions.

2.2.3 ANCHOR

ANCHORS was also developed by Marco Ribeiro (Ribeiro, Singh, & Guestrin, 2018) and is, again, a model-agnostic explanation approach based on if-then rules that are called *'anchors'*. These *'anchors'* are a set of feature conditions that act as high precision explainers created using reinforcement learning methods. This interpretability technique is not as computationally demanding as SHAP and is considered to have better generalisability than LIME.

There is a perception that Anchors provide a set of rules that are more easily understood by end users, although in this dissertation the analysis will be solely on the comparison of quantitative metrics. This will require a conversion of Anchor data into a numerical format for comparative statistical analysis.

2.2.4 DICE

DICE (Diverse Counterfactual Explanations) (Mothilal & Tan, 2020) is an XAI method developed to offer insights into machine learning model decisions by generating counterfactual explanations. In essence, a counterfactual explanation describes a minimal set of changes required to alter the model's prediction for a particular instance. For example, in a loan approval scenario, if an applicant was declined by a model, DICE could elucidate that increasing the annual income by a specific amount or improving the credit score by a few points would have led to an approval. This approach not only aids in understanding the model's behavior but also provides actionable feedback

to the end-users.

The strength of DICE lies in its ability to produce diverse counterfactuals that span the different dimensions of the feature space, enabling stakeholders to obtain a holistic view of the model's decision-making process(Nri, Jenkins, Paul, & Caruana, 2019).

Again, for the purposes of this paper, experiments with DiCE explanations will be converted into an input for a statistical analysis.

Chapter 3

Experiment design and methodology

3.1 Research Hypothesis for this Paper

Null Hypothesis:

It is not possible to quantify, and distinguish, the best interpretation framework to explain the reason for a specific (local) credit card fraud classification result using the following state-of-the-art techniques; SHAP, LIME, ANCHORS, and DICE.

Alternate Hypothesis:

IF a Neural Network algorithm is trained on a credit card transaction dataset for ML fraud detection, and SHAP, LIME, ANCHORS, and DICE interpretability frameworks are applied to individual model results

THEN a test for significance can be applied to the scores of each interpretability framework, against a predefined set of similarity metrics, to rank each explainer technique and demonstrate statistically which is best for explaining local credit card fraud classification results.

Section 3.2.2 of this paper provides the list of evaluation metrics to be used to measure the performance of each explainer technique in the experiments for this paper.

A Friedman Test will be applied across the four techniques using subsets of predictions, produced by the NN and EBM models, to rank the interpretability outputs for SHAP, LIME, ANCHORS, and DICE. A P-value output of this test of less than 0.05 will be considered sufficient evidence against the Null Hypothesis in favour of the Alternate.

The P-value in isolation is not sufficient for this research, as it will be necessary to determine the degree of separation of performance between the interpretability frameworks. It is an parallel objective to validate the assumption from Microsoft researchers that their EBM technique will score as well as *black box* models. A Wilcoxon signed-rank test will be applied pairwise on the interpretability techniques to measure the scale of difference, if any, in performance between each explainer method.

3.2 Design and Implementation

3.2.1 Research objectives and experimental activities

The aim of the research in this paper is to rank four selected interpretability frameworks (LIME, SHAP, Anchors, and DICE), using predefined similarity metrics, against the output from a Neural Network (NN) credit card fraud detection model and determine which one, if any, demonstrates the best overall performance.

The study will execute a number of research steps to build up a table of metrics for each explainer method and allow a statistical comparative analysis of the performance by each technique. The research focus is on explanations for fraud classification of individual transaction records – hence these experiments only consider local, post-hoc results.

The dataset for this study has been sourced from my employer, SymphonyAI, but relates to a product development cycle that ran from 2014 – 2018 by a subsidiary company (Norkom Technologies). The data was synthesised in 2013 from a number of US based credit card transaction sources and contains 25,128 rows, each one representing a credit card purchase. In this record set 15% of entries have been labelled as *'fraud'* by an analysis of which transactions were subsequently reported as fraudulent. The

data was used for product testing and demonstration purposes, but that particular product line was discontinued in 2019 and access has been granted to this, now redundant, dataset. The 2013 data generation process pulled in a significant amount of POS information, along with certain ETL attributes for use within the Norkom fraud application, resulting in a dataset of 380 columns.

The data has no missing values, and is free of any corruption in the data elements. The *'fraud'* label is a simple *'0'* or *'1'* binary value, *'1'* being used to represent that this given transaction record was deemed fraudulent. The model building exercise is thus a standard classification problem.

24K records will be used for model training, testing and refinement. 500 records will be set aside as 'unseen' date to produce a collection of 'explanations' for each individual records. This explanation dataset will be sub-divided into 20 batches for use in the research experiments to generate a table of numerical outputs against the following metrics (elaborated in Section 5.2 of this submission);

1. Identity
2. Stability
3. Separability
4. Similarity
5. Computational Efficiency

Figure 3.1 shows the diagrammatic view of experiment design for comparing explainability methods.

A very peripheral objective of this research is to assess the ease of use of cloud-based ML development options. Therefore, the experiments will be created and executed within an Kubeflow Studio integrated development environment (IDE). Kubeflow offers a Jupyter Notebook style interface, and the experiments will be written using Python 3.7. The resources assigned to each notebook kernel will be identical, particularly so that the *'Computational Efficiency'* metric can be compared accurately across all explainer techniques.

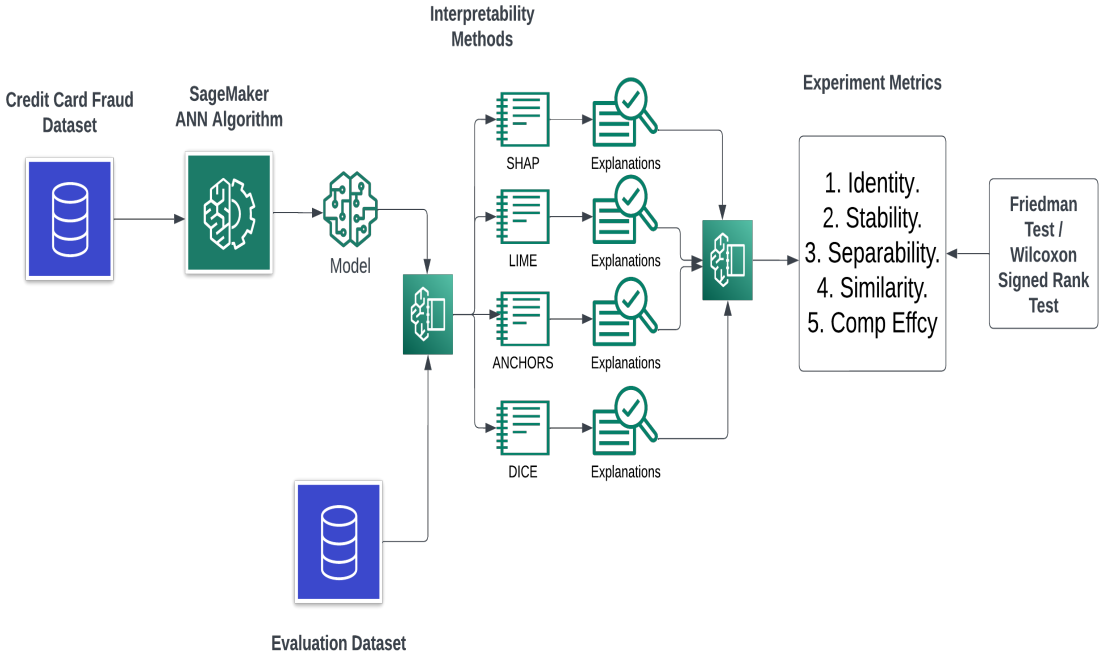

Figure 3.1: Overview of experiment design

The initial experiment steps will be to re-engineer the data prior to model creation. Credit Card fraud datasets often suffer from a severe imbalance of the target class (Priscilla & Prabha, 2020). However, in this case the fraudulent records in this research dataset comprise 15% of the entire data. While this is considerably more balanced than typical credit card fraud datasets, we will down sample the non-fraud records to create an even classification split. To simplify the process, and avoid adding any new synthetic data, a number of non-fraud records will be removed to that the remaining data set is 7K rows in size with a 50/50 breakdown of fraud v non-fraud. (Ribeiro et al., 2016) note that highly dimensional data can complicate the interpretability process, and it is generally desirable to focus on the key features for local explainer outputs.

Using the Kubeflow Notebooks, a basic classifier model can be created and used to identify and remove unnecessary highly correlated features. Canvas can also identify the top 20 features that contribute to the fraud classification results. Using this feature

list, the original dataset can be reduced to just these 20 column attributes and the fraud label column.

The model building exercise will begin with the reduced credit card fraud dataset. Using an inbuilt SageMaker ANN algorithm a fraud detection model will be built using a Training/Testing split of 80/20. This model will be providing predictions and explanations for three of the interpretability techniques. Taking comparative NN fraud detection experiments from (Sinanc et al., 2021) and (Anowar & Sadaoui, 2020), a target performance threshold of >= **0.85** and >= **0.90** will apply for **F1** and **Recall** respectively. This will ensure that a performant NN model has been created prior to the measurements of the results from the experiments on the separate interpretability frameworks.

The 500 credit card transaction records are processed by both models to produce two sets of predictions.

This set of data is split into 20 sub-groups and sets of explanations are generated and scored for each batch of data.

The SHAP, LIME, ANCHORS, and DICE explainability techniques are used to generate the explanations from the ANN model.

The form of the research is to gather knowledge from the numerical results of the experiments and determine if the frameworks can be clearly ranked in terms of overall performance by the applied metrics. This approach follows some of the concepts in measuring similarity performance for explainability techniques as elaborated by (ElShawi et al., 2020). This will be a deductive approach to test the assumption that one particular interpretability frameworks can be shown, through statistical significance testing on the numerical outputs of each experiment, to generate the best local explanations for a credit card fraud classification result. Although the experiments of (Evans et al., 2019) focused on global explanations, their experiments used a Friedman test to collate p-values into a correlation matrix and while the metrics used are different to the ones proposed in this paper this is a general approach that will be emulated in this dissertation.

3.2.2 Evaluation of designed solution with performance metrics (and statistical tests)

The explainability metrics proposed below extend the framework comparison research conducted by (ElShawi et al., 2020), but transfers the domain from healthcare analysis to fraud detection. (ElShawi et al., 2020) was in turn influenced by papers from (Honegger, 2018) and (Guidotti et al., 2019).

1. Identity. A measure of how much identical instances have identical explanations. For every two instances in the testing data if the distance between features is equal to zero, then the distance between the explanations should be equal to zero.

2. Stability. Instances belonging to the same class have comparable explanations. K-means clustering is applied to explanations for each instance in test data. Measure the number of explanations in both clusters (fraud/non-fraud) that match predicted class for instance from NN model.

3. Separability. Dissimilar instances must have dissimilar explanations. Take subset of test data and determine for each individual instance the number of duplicate explanations in entire subset, if any.

4. Similarity. This metric captures the assumption that the more similar the instances to be explained, the closer their explanation should be (and vice versa). Cluster test data instances into Fraud/non-Fraud clusters. Normalise explanations and calculate Euclidean distances between instances in both clusters. Smaller mean pairwise distance = better explainability framework metric.

5. Computational Efficiency. Average time taken, in seconds, by the interpretability framework to output a set of explanations. (Similar Cloud environments are applied to all experiments).

A metric such as *'Computational Efficiency'* could be considered unrelated to a measure of explainability, but this research proposal contends that it is important to

consider in terms of feasibility across XAI methods. Computational time can be a bottleneck in generating explanations and may have an impact on the commercial viability of an explainability process in a commercial credit card fraud detection application

A Friedman test will be run to determine if evidence exists that there is a difference in performance between SHAP, LIME, ANCHORS, and DICE in terms of explaining local credit card fraud classification results. The research assumption will be that a calculated P-value of less than 0.05 implies that a given technique can be ranked higher than the others.

A subsequent Wilcoxon signed-rank test would be run on each pair of interpretability techniques to measure of the degrees of separation.

A P-value of greater than 0.05 will provide evidence that the explainer techniques examined in this paper do not show significant differences in performance, supporting the Null Hypothesis in the research question.

Chapter 4

Results, evaluation and discussion

4.1 Results...

The output from the experiments produces a matrix of metrics results in the following format 4.1.

Table 4.1: Table of XAI Metrics Results

	SHAP	**LIME**	**ANCHORS**	**DiCE**
Identity	0.25	0.77	0.67	0.87
Stability	0.2	0.33	0.32	0.73
Separability	0.41	0.52	0.41	0.27
Similarity	0.44	0.05	0.99	0.07
Computational Efficiency	0.94	0.25	0.21	0.11

4.2 Evaluation...

4.3 Discussion...

Chapter 5

Conclusion

5.1 Research Overview

5.2 Problem Definition

5.3 Design/Experimentation, Evaluation & Results

5.4 Contributions and impact

5.5 Future Work & recommendations

References

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (xai). *IEEE Access*, *6*(52), 138–160. doi: 10.1109/access.2018.2870052

Ajitha, E., Sneha, S., Makesh, S., & Jaspin, K. (2023). A comparative analysis of credit card fraud detection with machine learning algorithms and convolutional neural network. In *2023 international conference on advances in computing, communication and applied informatics (accai)* (p. 1-8). doi: 10.1109/ACCAI58221.2023.10200905

Alvarez-Melis, D., & Jaakkola, T. (2018). On the robustness of interpretability methods. *2018 ICML Workshop on Human Interpretability in Machine Learning*. doi: 10.48550/arXiv.1806.08049

Anowar, F., & Sadaoui, S. (2020). Incremental neural-network learning for big fraud data. *2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, *1*(1), 1–4. doi: 10.1109/smc42975.2020.9283136

Aurna, N. F., Hossain, M. D., Taenaka, Y., & Kadobayashi, Y. (2023). Federated learning-based credit card fraud detection: Performance analysis with sampling methods and deep learning algorithms. In *2023 ieee international conference on cyber security and resilience (csr)* (p. 180-186). doi: 10.1109/CSR57506.2023.10224978

Batageri, A., & Kumar, S. (2021). Credit card fraud detection using artificial neural network. *Global Transitions Proceedings*, *2*(1), 35–41. doi: 10.1016/j.gltp.2021.01.006

Carvalho, D. V., Pereira, E. M., & Cardoso, J. S. (2019, Jul). Machine learning interpretability: A survey on methods and metrics. *Electronics*, *8*(8). doi: 10.3390/electronics8080832

Chromik, M., .Eiband, M., Buchner, F., Krüger, A., & Butz, A. (2021). I think i get your point, ai! the illusion of explanatory depth in explainable ai. *26th International Conference on Intelligent User Interfaces*. doi: 10.1145/3397481.3450644

Dal Pozzolo, A., et al. (2014). Learned lessons in credit card fraud detection from a practitioner perspective. *Expert Systems with Applications*, *41*(10), 4915–4928. doi: 10.1016/j.eswa.2014.02.026

Darias, J. M., Caro-Martínez, M., Díaz-Agudo, B., & Recio-Garcia, J. A. (2022, Aug). Using case-based reasoning for capturing expert knowledge on explanation methods. *Case-Based Reasoning Research and Development*, *13405*, 3–17. doi: 10.1007/978-3-031-14923-8_1

ElShawi, R., Sherif, Y., Al-Mallah, M., & Sakr, S. (2020, Aug). Interpretability in healthcare: A comparative study of local machine learning interpretability techniques. *Computational Intelligence*, *37*(4), 1633–1650. doi: 10.1111/coin.12410

Evans, B. P., Xue, B., & Zhang, M. (2019, Jul). What's inside the black-box? *Proceedings of the Genetic and Evolutionary Computation Conference*. doi: 10.1145/3321707.3321726

Guidotti, R., Monreale, A., Giannotti, F., Pedreschi, D., Ruggieri, S., & Turini, F. (2019, Dec). Factual and counterfactual explanations for black box decision making. *IEEE Intelligent Systems*, *34*(6), 14–23. doi: 10.1109/mis.2019.2957223

Hanafy, M., & Ming, R. (2022). Classification of the insureds using integrated machine learning algorithms: A comparative study. *Applied Artificial Intelligence*, *36*. doi: 10.1080/08839514.2021.2020489

Honegger, M. (2018, Aug). *Shedding light on black box machine learning algorithms: Development of an axiomatic framework to assess the quality of methods that explain*

individual predictions. Karlsruhe Institute of Technology. Retrieved from `https://arxiv.org/abs/1808.05054v1`

Ignatiev, A. (2020, Jul). Towards trustable explainable ai. *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, 5154–5158. doi: 10.24963/ijcai.2020/726

Jacob, V., Song, F., Stiegler, A., Rad, B., Diao, Y., & Tatbul, N. (2021). Exathlon: A benchmark for explainable anomaly detection over time series. *Proceedings of the VLDB Endowment*, *14*(11), 2613–2626. doi: 10.14778/3476249.3476307

Jesus, S., Belém, C., Balayan, V., Bento, J., Saleiro, P., Bizarro, P., & Gama, J. (2021, Mar). How can i choose an explainer? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. doi: 10.1145/3442188.3445941

Kaur, H., Nori, H., Jenkins, S., Caruana, R., Wallach, H., & Wortman-Vaughan, J. (2020, Apr). Interpreting interpretability: Understanding data scientists' use of interpretability tools for machine learning. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–14. doi: 10.1145/3313831.3376219

Lakkaraju, H., Bach, S. H., & Leskovec, J. (2016, Aug). Interpretable decision sets. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1675–1684. doi: 10.1145/2939672.2939874

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in neural information processing systems 30 (nips 2017)* (Vol. 30). NeurIPS Proceedings.

Marcilio, W. E., & Eler, D. M. (2020, Nov). From explanations to feature selection: Assessing shap values as feature selection mechanism. *2020 33rd SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, 340–347. doi: 10.1109/sibgrapi51738.2020.00053

Martínez, M., Nadj, M., Langner, M., Toreini, P., & Maedche, A. (2023). Does this explanation help? designing local model-agnostic explanation representations

and an experimental evaluation using eye-tracking technology. *ACM Transactions on Interactive Intelligent Systems*. doi: 110.1145/3607145

Moreira, C., Chou, Y., Velmurugan, M., Ouyang, C., Sindhgatta, R., & Bruza, P. (2020). Linda-bn: An interpretable probabilistic approach for demystifying black-box predictive models. *Decision Support Systems*, *150*. doi: 10.1016/j.dss.2021.113561

Mothilal, S. A., R. K., & Tan, C. (2020). Explaining machine learning classifiers through diverse counterfactual explanations. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. doi: 10.48550/arXiv.1806.08049

Nascita, A., Montieri, A., G.Aceto, Ciuonzo, D., Persico, V., & Pescape, A. (2021). Unveiling mimetic: Interpreting deep learning traffic classifiers via xai techniques. *2021 IEEE International Conference on Cyber Security and Resilience (CSR)*, 455–460. doi: 10.1109/csr51186.2021.9527948

Nesvijevskaia, A., Ouillade, S., Guilmin, P., & Zucker, J. (2021). The accuracy versus interpretability trade-off in fraud detection model. *Data and Policy*, *3*. doi: 10.1017/dap.2021.3

Nguyen, M., Bouaziz, A., Valdes, V., Rosa-Cavalli, A., Mallouli, W., & Montes-DeOca, E. (2023). A deep learning anomaly detection framework with explainability and robustness. *Proceedings of the 18th International Conference on Availability, Reliability and Security.*. doi: 10.1145/3600160.3605052

Nri, H., Jenkins, S., Paul, K., & Caruana, R. (2019). Interpretml: A unified framework for machine learning interpretability. *arXiv preprint arXiv:1909.09223*.

Priscilla, C., & Prabha, D. (2020). Influence of optimizing xgboost to handle class imbalance in credit card fraud detection. *2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT)*, 1309–1315. doi: 10.1109/icssit48917.2020.9214206

Psychoula, I., Gutmann, A., Mainali, P., Lee, S. H., Dunphy, P., & Petitcolas, F. (2021). Explainable machine learning for fraud detection. *Computer*, *54*(10), 49–59. doi: 10.1109/mc.2021.3081249

Ras, G., Xie, N., Gerven, M. V., & Doran, D. (2022). Explainable deep learning: A field guide for the uninitiated. *Journal of Artificial Intelligence Research*, *73*, 329–397. doi: 10.1613/jair.1.13200

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016, Aug). "why should i trust you?" explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144. doi: 10.1145/2939672.2939778

Ribeiro, M. T., Singh, S., & Guestrin, C. (2018, Feb). Anchors: High-precision model-agnostic explanations. *Proceedings of the AAAI Conference on Artificial Intelligence*, *32*(1). doi: 10.1609/aaai.v32i1.11491

Schlegel, U., Arnout, H., El-Assady, M., Oelke, D., & Keim, D. (2019). Towards a rigorous evaluation of xai methods on time series. *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*.. doi: 10.1109/iccvw.2019.00516

Sharma, A., & Bathla, N. (2020, Aug).

Review on credit card fraud detection and classification by Machine Learning and Data Mining approaches, *6*(4), 687–692.

Sharma, P., & Priyanka, S. (2020, Jun). Credit card fraud detection using deep learning based on neural network and auto encoder. *International Journal of Engineering and Advanced Technology*, *9*(5), 1140–1143. doi: 10.35940/ijeat.e9934.069520

Sinanc, D., Demirezen, U., & Sağıroğlu, (2021). Explainable credit card fraud detection with image conversion. *ADCAIJ: Advances in Distributed Computing and Artificial Intelligence Journal*, *10*(1), 63–76. doi: 10.14201/adcaij20211016376

Sullivan, R., & Longo, L. (2023). Explaining deep q-learning experience replay with shapley additive explanations. *Machine Learning and Knowledge Extraction*, *5*(4), 1433–1455. doi: 10.48550/arXiv.1806.08049

T.Y.Wu, & Y.T.Wang. (2021). Locally interpretable one-class anomaly detection for credit card fraud detection. *2021 International Conference on Technologies and Applications of Artificial Intelligence (TAAI)*. doi: 10.1109/taai54685.2021.00014

Vilone, G., & Longo, L. (2021a, May). Notions of explainability and evaluation approaches for explainable artificial intelligence. *Information Fusion*, *76*, 89–106. doi: 10.1016/j.inffus.2021.05.009

Vilone, G., & Longo, L. (2021b). A quantitative evaluation of global, rule-based explanations of post-hoc, model agnostic methods. *Frontiers in Artificial Intelligence*, *4*. doi: 10.3389/frai.2021.717899

Vouros, G. (2022). Explainable deep reinforcement learning: State of the art and challenges. *ACM Computing Surveys*, *55*(5), 1–39. doi: 10.1145/3527448

Appendix A

Additional content

These should contain supplementary material that is not necessary in order for the reader to follow the argument. For example, the text of a questionnaire, detailed UML diagrams, or a complete Software Requirement Specification should be placed in an Appendix. It is not considered necessary to include code, but you may do so by including a link within your dissertation PDF.