Research Design & Proposal Writing (Full Proposal): Explaining Credit Card Fraud Decisions in ML: An Analysis of XAI Methods

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

Student Name Ciaran Finnegan - D21124026
TU060 - MSc in Data Science - Technological University Dublin

1 Background, Context and Scope

Credit card fraud costs the Financial Services industry billions of Euros of loss each year. 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 fraud classifications that they make.

Ignatiev (2020) focuses on the need for Explainable Artificial Intelligence (XAI) to be *trustable*, while Carvalho et al. (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. Companies operating in the area of financial crime software, such as SymphonyAI and Actimise, already sell ML based software to detect credit card fraud but generally rely on only one explainer technique, such as SHAP values

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

2 Problem Description

2.1 Approaches to solve the 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'*.

Metrics for all four explainer techniques will be collated and subjected to a statistical test for significance. Is one explainer better than another and if so, how great is that difference?

The proposed experiments in this paper are based on a similar study into measuring interpretability methods on healthcare datasets that classified mortality predictions (ElShawi et al., 2020). A key assumption is that this research approach will translate into the domain of credit card fraud.

If 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.

Experiments are being specifically limited to four post hoc and local interpretability frameworks in order to build on related research papers by Ribeiro et al. (2016) and Guidotti et al. (2019). Only local explanations on specific credit card transactions are being considered – global explainability on the overall model is not in scope. Deliberately, there is no human assessment of the explanations as this will be a purely programmatic and arithmetic exercise.

2.2 Gaps in Research

The literature review (to date) for this dissertation proposal 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 et al. (2021). Despite this acknowledgement, in this research domain there appears to be no cast iron process to establish this 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.

XAI research in the domain of healthcare is more commonplace (Marcilio & Eler, 2020) (Lakkaraju et al., 2016) and often involve experiments with clearly objective recommendations (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 (Jesus et al., 2021). This dissertation will follow in the steps of earlier research that use experiments with quantifiable metrics (Darias et al., 2022) and tests for statistical significance (Evans et al., 2019).

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.

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.

2.3 State Of The Art Approaches

This section of the document describes the local interpretability techniques that will form the basis of the experiments in this dissertation proposal.

2.3.1 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).

2.3.2 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.

2.3.3 ANCHOR

ANCHORS was also developed by Marco Ribeiro (Ribeiro et al., 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.

2.3.4 DICE (Diverse Counterfactual Explanations)

DICE (Diverse Counterfactual Explanations) is an XAI (Explainable Artificial Intelligence) 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.

3 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.

4 Hypothesis

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 5.2 of this proposal 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.

5 Design and Implementation

5.1 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 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 Amazon AWS SageMaker Studio integrated development environment (IDE). SageMaker 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.

The initial experiment steps will be to re-engineer the data prior to model creation. The fraudulent records comprise 15% of the entire data, and 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 Amazon SageMaker Studio Canvas application, 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

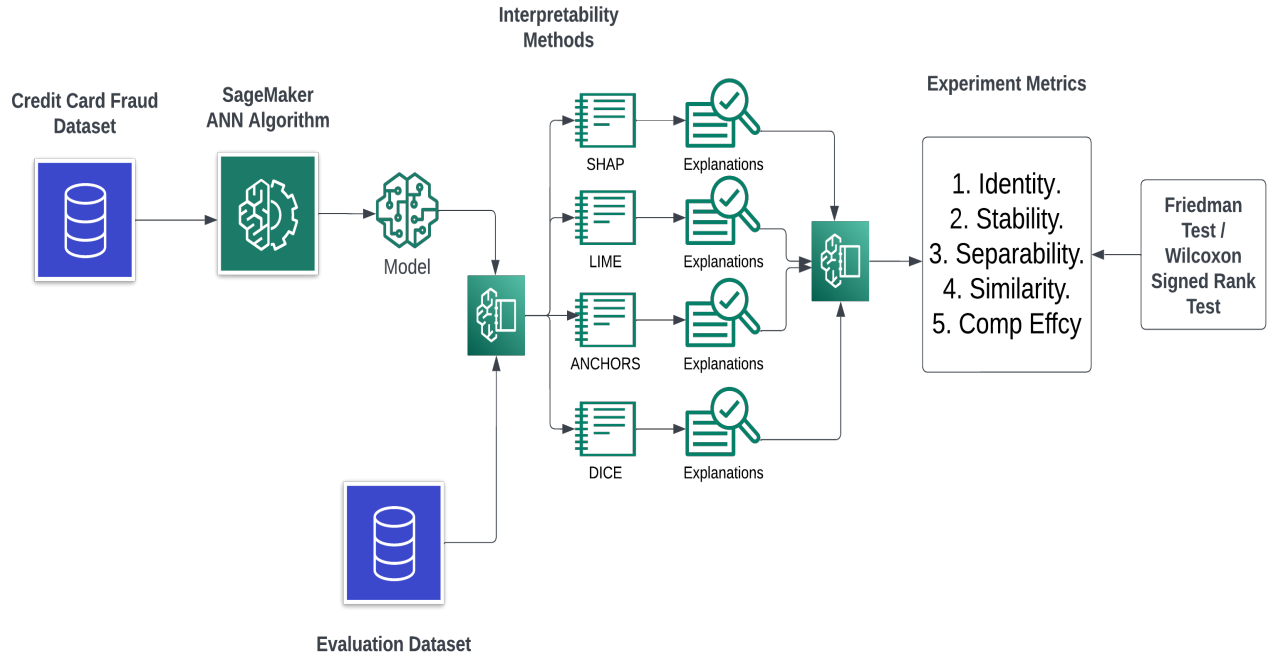

Figure 1: *Overview of experiment design*

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.

5.2 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 matric 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.

6 Activities

Following an AGILE software development mindset, activities are broken into a series of *'User Stories'*. This reflects the intention that each activity task has a clearly defined goal at the outset, and a measure of success at the end.

User Story 1

Although preparation for this submission involved a review of 45+ papers in the research field of XAI, further research into similar comparative experiments for explainer methods will be necessary as a starting activity. This initial period (User Story 1) will also be used to set up an AWS SageMaker account and run 3+ trial Python notebook exercises.

User Story 2

Data Preparation takes place in week 3 (User Story 2) to reduce the feature set and balance the *'fraud'* and *'non-fraud'* classes.

User Story 3 + 4

Week 4 is the model building activity (User Story 3). One model will be built with a SageMaker ANN algorithm, the other will be created using libraries imported from the InterpretML GitHub repository. User Story 4 will focus on the most complex and lengthy period of dissertation activity; producing the output scores for SHAP, LIME, and ANCHORS explainability methods.

User Story 5 + 6

User Story 5 is an interim step to carry out significance testing on the partial results gathered so far. This work leads onto User Story 6, which is the creation of an interim report. This will allow for supervisor feedback at a point when approximately 60% of the dissertation work should be complete.

User Story 5 + 6

The next experiment introduces the Counterfactual *'DICE'* explainers in User Story 7. Metrics are updated and tests are re-run in User Story 8 to validate if a statistical differences exist between the interpretability methods.

Findings from the experiments are written into the final document, along with whatever additional updates are appropriate during User Story 9 to allow submission of the finished paper.

Figure 2 and 3 show the Gantt Charts with timescales for the research activities in this dissertation.

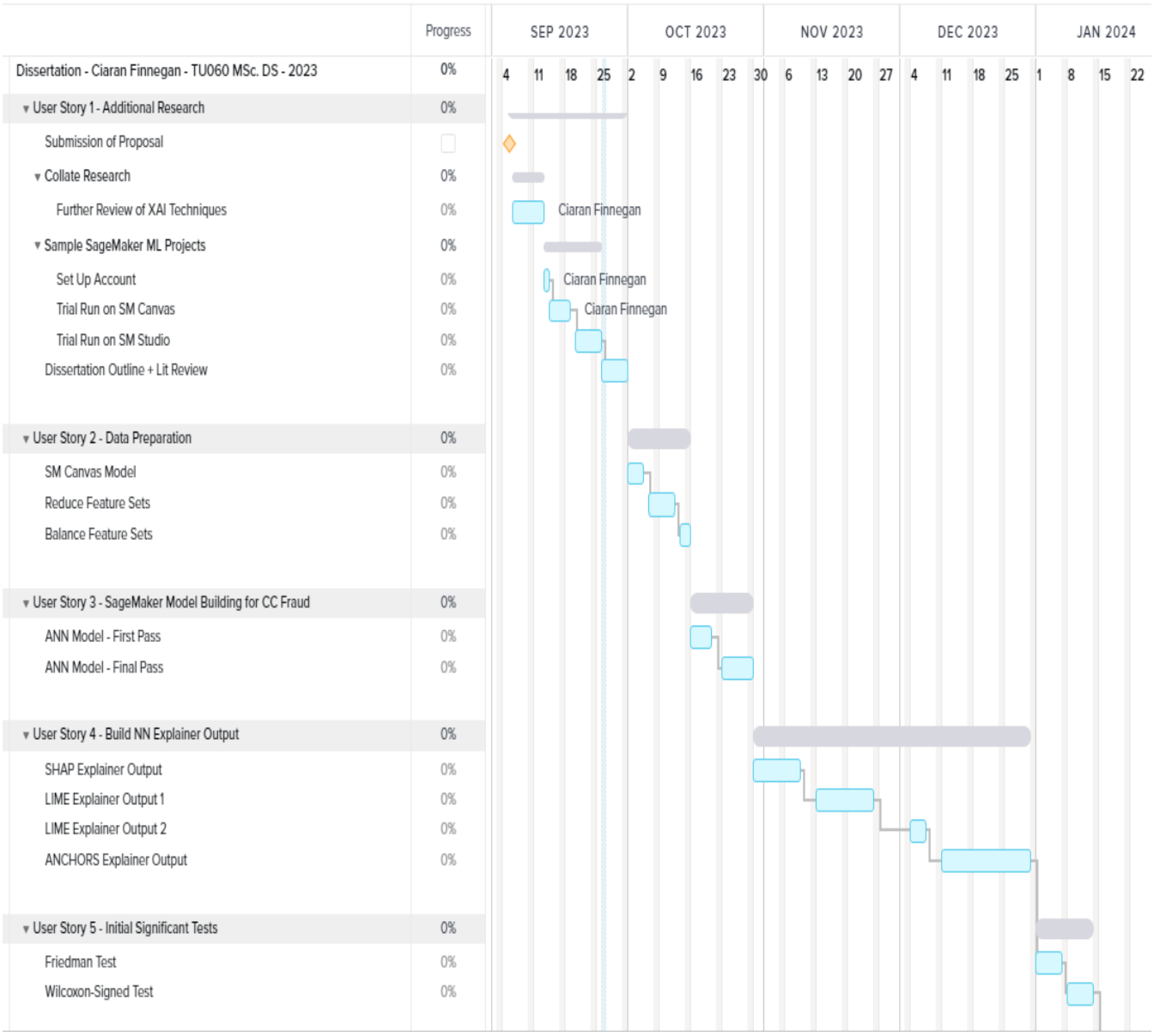

Figure 2: *Gantt Chart - September 2023 - January 2024*

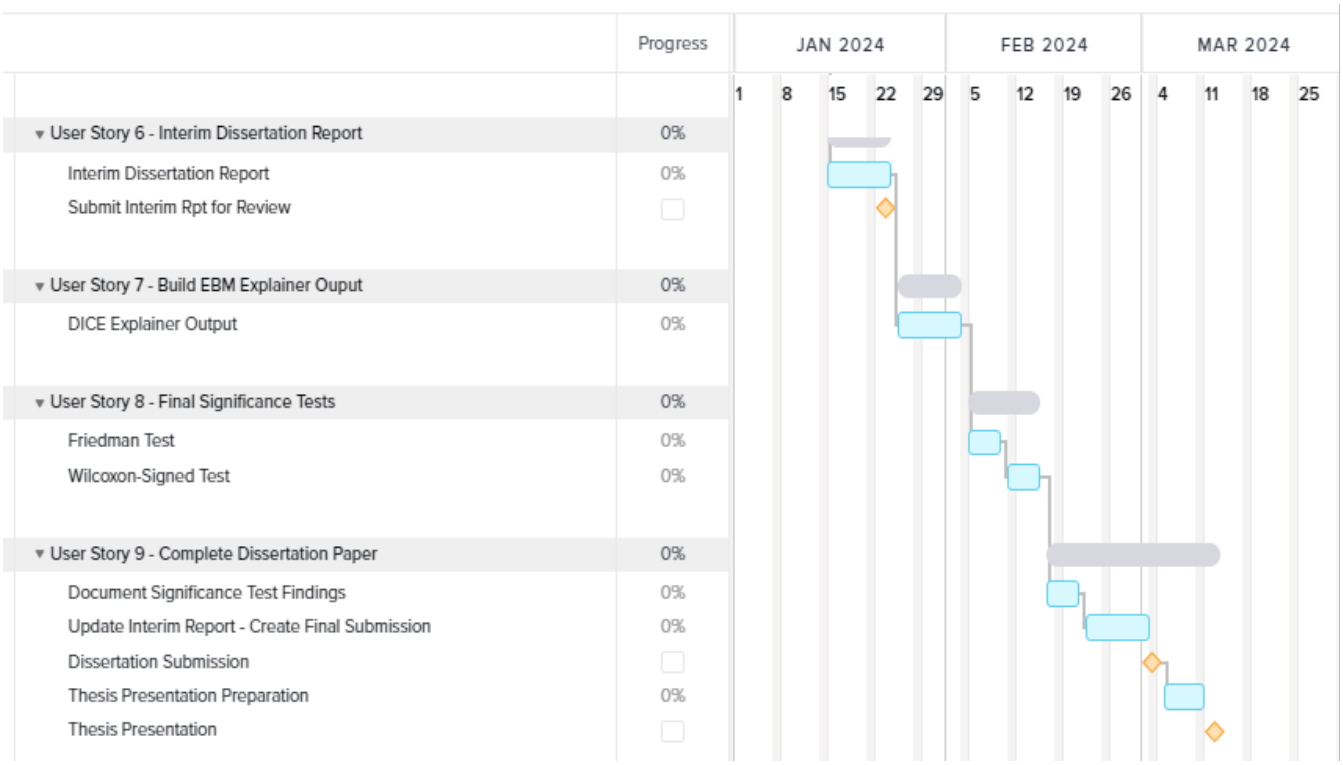

Figure 3: *Gantt Chart - January 2023 - March 2024*

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

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

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

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

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

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

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

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

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

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

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

7 Appendices

Assignment 1 and Assignment 2 submissions (PDF) are reproduced here in this section.

7.1 Assignment 1

Research Design + Proposal Writing (CA1)

- Research Question
- Hypothesis
- Preliminary Design

CIARAN FINNEGAN

TU060 YR. 2 DATA SCIENCE

STD NO: D21124026

Domain, scope, assumptions, limitations and delimitations of research - ACM 2012

DOMAIN:

A: *Applied Computing → Electronic Commerce → Digital Cash* (Anowar & Sadaoui, 2020)

B: Social and Professional Topics → Computing / Technology Policy → Computer Crime → Financial Crime (Sharma & Priyanka, 2020; Psychoula et al., 2021)

C: *Applied Computing → Computer Forensics → Investigation Techniques* (Sharma & Bathla, 2020)

D: *Computing Methodologies → Machine Learning → Machine Learning Approaches → Neural Networks* (Batageri & Kumar, 2021; Anowar & Sadaoui, 2020)

E: *Computing Methodologies → Artificial Intelligence → Knowledge Representation and Reasoning → Causal Reasoning and Diagnostics* (Vilone & Longo, 2021; Sinanc et al., 2021; Psychoula et al., 2021; Adadi & Berrada, 2018; Lundberg & Lee 2017)

SCOPE : Using widely available cloud-based technologies, develop a small scale online ML application for credit card fraud detection; that shows a Neural Network algorithm can provide an Explainable AI (XAI) method to interpret why a record is classified as fraud.

ASSUMPTIONS: 15% of the records in the dissertation dataset are labelled as 'fraud', therefore it will not be necessary to pre-process the data with any synthetic data generation, or over/under sampling techniques; the modelling and production deployment options, which include XAI outputs, can all be developed on Amazon SageMaker; the production model will deliver a ~3 second response, which includes the fraud classification result and explanation.

LIMITATIONS : SHAP (**SH**apley **A**dditive ex**P**lanations) is a prominent method to explain ML classifications (fraud detection in this dissertation), but as it requires the use of a 'background data set' to infer its values for feature ranking it may be necessary to avoid the use of the full dataset for performance reasons (with possible impact on the accuracy of explanations).

DELIMITATIONS: Dissertation research is limited to US Credit Card Fraud transactions as this is the best available internal dataset from within my FinTech company (250K records); as this is a labelled dataset, only a supervised ML approach is being considered to build the NN model; the dataset contains 300+ features, so feature selection will be applied, in early iterations of the ML workflow process, to focus on the columns providing the most understandable explanations.

Gaps in the literature and research question

Gaps: Data Availability and Handling Data Imbalance

1. Due to data confidentiality concerns, there are still relatively few historical credit card fraud datasets upon which to conduct ML experiments for any aspect of fraud detection, XAI or otherwise. This is a limitation noted in research conducted by Dal Pozzolo et al. (2014) and results in a small group of datasets frequently being re-used in multiple papers. Fortunately, I have access to a 'new' internal company compiled dataset of 250k credit card fraud records that may avoid potential bias in other datasets, and ideally has the detail to feed into meaningful XAI outputs.

2. Credit Card Fraud datasets tend to be heavily imbalanced. There are differences in the literature on how to take concrete steps to tackle this problem and avoid model bias. Priscilla & Prabha (2020) propose that resampling techniques themselves could be distorting credit card fraud data, which will impact on downstream results, including XAI outputs. In the dataset proposed for this dissertation, 15% of the records represent fraudulent transactions. Therefore, I will avoid resampling as a pre-processing step.

Gaps: How exactly does a researcher measure and display 'explainability' in Explainable Artificial Intelligence Research?

1. In their research experiments with the LIME (**L**ocal **I**nterpretable **M**odel-agnostic **E**xplanations) algorithm, Ribeiro et al. (2016) describe how users can have a *trust* issue with ML models, like NN, that 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, and there is no cast iron process to ensure this trustworthiness. This dissertation hopes to build on this body of work in the area of credit card fraud detection, and attempt to address the gaps listed here below.

2. 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 (2021) 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, there is no well established output framework for explaining credit card fraud classification through 'black-box' models.

3. The 'If-Then' style of rules could be an alternate XAI output option to be chosen for this dissertation. Vilone & Longo (2021) also assert that there is still relatively little research that objectively assesses this approach with quantitative metrics, thus allowing it to be benchmarked against other XAI methods.

4. Psychoula et al (2021) state that the runtime implications of XAI output (explanations) on real-time systems, fraud or otherwise, has had relatively little research focus to date. This dissertation aims to build a workable real-time interface to a credit card fraud detection ML production model, so a ~3 second response time for results and explanations will be part of the success criteria. Early prototyping in the dissertation effort will attempt to capture and address any such issues as early as possible.

Research Question: Is it possible to clearly explain to a financial auditor/investigator, in 'real-time', the explicit reasons why the attribute values of a given credit card transaction resulted in a Neural Network ML model classifying that record as fraudulent?

Hypothesis

Null Hypothesis

The conventional view is that for most observers the working of Neural Network algorithms are a 'black-box' process, and it is not possible to easily understand, and audit, why a given end result, such as a classification category, has been generated.

Alternate Hypothesis

IF I train a Neural Network algorithm for use in an ML process built, using cloud-based technology, for credit card fraud detection,
THEN the real-time model output will demonstrate a high ***Recall*** value, and for a specific 'local' instance record will contain the top 10 most important features, as ranked by both SHAP and LIME outputs, that drove the classification result.

Feasibility of the Study – Sequence of Tasks Planned

			Weeks	20
Task	**Description**	**Additional Comment**	**Duration**	**Remaining**
1	Working ***prototype/baseline*** logistic regression model trained/deployed in a Cloud ML workspace.	Credit card fraud dataset is already in place. Outputs, including feature importance, assessed only within the cloud ML workspace.	1.5	18.5
2	Feature selection on dataset to focus on 40+ most relevant, and explainable, attributes.	Re-run baseline model in cloud ML workspace.	1	17.5
3	Select appropriate NN algorithm for explainability project and train/deploy new model.	Ensure F1 and recall performance criteria met. Run in cloud workspace.	1	16.5
4	Generate SHAP values from NN model for key features explaining fraud classification results, and document.	Compare against feature importance from logistic regression baseline model.	2	14.5
5	Generate LIME explanations from NN model for key features explaining fraud classification results, and document.	As above.	2	12.5
6	Document Interim findings on hypothesis testing objectives.		2	10.5
7	Build external hosted UI interface to allow real time input of 'unseen' fraud data.	Return Classification result to external UI.	2	8.5
8	Augment UI interface with graphical display of model explanations.	Demonstrate if application can present hypothesis proof (or not).	1	7.5
9	Retune model and model explanations output.	Begin final documentation.	1.5	6
10	Refine UI.	Document description of UI.	1	5
11	Complete dissertation documentation.		4	1
12	Contingency		1	0

Bibliography

1. Adadi, A., & Berrada, M. (2018). Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, *6*(52), 138–160. https://doi.org/10.1109/access.2018.2870052

2. 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. https://doi.org/10.1109/smc42975.2020.9283136

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

4. Dal Pozzolo, A., Caelen, O., Le Borgne, Y.-A., Waterschoot, S., & Bontempi, G. (2014). Learned lessons in credit card fraud detection from a practitioner perspective. *Expert Systems with Applications*, *41*(10), 4915–4928. https://doi.org/10.1016/j.eswa.2014.02.026

5. 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). essay, NeurIPS Proceedings.

6. Priscilla, C. V., & Prabha, D. P. (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. https://doi.org/10.1109/icssit48917.2020.9214206

7. 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. https://doi.org/10.1109/mc.2021.3081249

Bibliography

8. 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. https://doi.org/10.1109/mc.2021.3081249

9. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "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. https://doi.org/10.1145/2939672.2939778

10. Sharma, A., & Bathla, N. (2020). *Review on Credit Card Fraud Detection and Classification by Machine Learning and Data Mining Approaches*, *6*(4), 687–692. Retrieved from https://www.semanticscholar.org/paper/Review-on-credit-card-fraud-detection-and-by-and-Sharma-Bathla/b6c839cadb4c6281a934a8788fec93d5482e6af4.

11. Sharma, P., & Priyanka, S. (2020). 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. https://doi.org/10.35940/ijeat.e9934.069520

12. Sinanc, D., Demirezen, U., & Sağıroğlu, Ş. (2021). Explainable Credit Card Fraud Detection with Image Conversion. *ADCAIJ: Advances in Distributed Computing and Artificial Intelligence Journal*, *10*(1), 63–76. https://doi.org/10.14201/adcaij20211016376 A new explainable artificial intelligence approach is ... presented. In this way, feature relationships that have a dominant effect on fraud detection are revealed.

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

14. Vilone, G., & Longo, L. (2021). Classification of explainable artificial intelligence methods through their output formats. *Machine Learning and Knowledge Extraction*, *3*(3), 615–661. https://doi.org/10.3390/make3030032

Performance Metrics for Experiments

Explainability Metrics;

1. Create a baseline with a logistic regression classifier that has been modelled against the dissertation data. This baseline model will measure individual feature importance, through coefficient weights. The NN model output (result and explanation), will have associated SHAP and LIME values indicating that model's list of important features. The performance expectation is that both models match at least 70% - 80% of the same key attribute values. This metric is based on general outputs of credit card fraud experimental data from Psychoula et al. (2021).

2. Compare real time response of production NN model using SHAP v. LIME. Determine if a subsampled background set for SHAP can match LIME for speed and accuracy of explanation (both will target ~3 secs to respond with values). Again this metric follows related experiment data in the Psychoula et al. (2021) paper. The purpose is to demonstrate that the model can deliver accurate classification explanations in an acceptable timeframe for both algorithms.

Metrics to apply to any meaningful credit card fraud model;

1. **F1** is a better score for fraud detection problems, as opposed to simple accuracy, because of the uneven class distribution seen in many credit card datasets. This score takes the numbers of false positives and false negatives into a weighted average. Taking comparative NN fraud detection experiments from Sinac et al. (2021), a target threshold of **>= 0.85** will apply to the experiments in this dissertation.

2. In conjunction with F1, **Recall** will be used as a measure as this reflects the model's ability to detect positive samples, which is important in any credit card fraud detection system. Using experiment metrics applied by Anowar & Sadaoui (2020), a target Recall value will be set of **>= 0.9**.

3. As above, a response time from the production model of < 4 secs is expected, including both the classification result and a 'local' interpretable output explaining the reason for any 'fraud' result. The dissertation app should mimic the general performance expectation of any Web app.

Fw: feedback 1st assignment

D21124026 Ciaran Finnegan

To: ciaran@feefinnegan.com; ciaran.finnegan@netreveal.ai

From: Luca Longo <Luca.Longo@TUDublin.ie>
Sent: Saturday, November 12, 2022 5:20 PM
To: D21124026 Ciaran Finnegan <D21124026@mytudublin.ie>
Subject: feedback 1st assignment

1	S,A,L	2	2.00	2.00	2	1	1	11
domain - 3	wrong (D=domain, S=scope, L=limitations, D=delimitations)	gaps - 3	research question - 3	hypothesis - 5	feasibility - 3	bibliography - 2	Evaluation Metrics - 1	mark 1st assignment

Dr. Luca Longo
Artificial Intelligence & Cognitive Load Research Lab (AICL-lab)

7.2 Assignment 2

Research Design + Proposal Writing (CA2)

- Research Question
- Hypothesis
- Refined Design

CIARAN FINNEGAN

TU060 YR. 2 DATA SCIENCE

STD NO: D21124026

Domain, scope, assumptions, limitations and delimitations of research - ACM 2012

DOMAIN:

A: *Applied Computing → Electronic Commerce → Digital Cash* (Anowar & Sadaoui, 2020)

B: Social and Professional Topics → Computing / Technology Policy → Computer Crime → Financial Crime (Dal Pozzolo et al,2014, Sharma & Priyanka, 2020; Psychoula et al., 2021)

C: *Applied Computing → Computer Forensics → Investigation Techniques* (Sharma & Bathla, 2020; Honegger, 2018; Ribeiro et al., 2016)

D: *Computing Methodologies → Machine Learning → Machine Learning Approaches → Neural Networks* (Batageri & Kumar, 2021; Anowar & Sadaoui, 2020)

E: *Computing Methodologies → Artificial Intelligence → Knowledge Representation and Reasoning → Causal Reasoning and Diagnostics* (Vilone & Longo, 2021; Sinanc et al., 2021; Psychoula et al., 2021; Adadi & Berrada, 2018; Lundberg and Lee 2017; Guidotti et al., 2019; ElShawi et al., 2020)

SCOPE : To assess how post hoc, local interpretability frameworks can be evaluated to improve the quality of explanation for neural network models generating credit card fraud classifications in a commercial application.

ASSUMPTIONS : 15% of the records in the dissertation dataset are labelled as 'fraud', therefore it will not be necessary to pre-process the data with any synthetic data generation, or over/under sampling techniques; the modelling and production deployment options, which include XAI outputs, can all be developed on Amazon SageMaker; the production model will deliver a ~4 second response, which includes the fraud classification result and explanation.

LIMITATIONS : This research must work within environmental constraints that are commercially viable, hence the time taken to generate explanations is a factor and may impact on experiments, particularly using SHAP values; cloud-based environments will be deployed but the use of extensive GPU processing is expensive and beyond what can be afforded for the experiments in this dissertation.

DELIMITATIONS : Experiments are being specifically limited to five post hoc and local interpretability frameworks; LIME, SHAP, Anchors, LORE, and InterpretML (Microsoft) in order to build on research by Guidotti et al., (2019), ElShawi et al, (2020), Ribeiro et al., (2016); Only local explanations on specific credit card transactions are being considered – global explainability on the overall model is not in scope.

Gaps in the literature and research question

Gaps: Data Availability and Handling Data Imbalance

1. Due to data confidentiality concerns, there are still relatively few historical credit card fraud datasets upon which to conduct ML experiments for any aspect of fraud detection, XAI or otherwise. This is a limitation noted in research conducted by Dal Pozzolo et al. (2014) and results in a small group of datasets frequently being re-used in multiple papers such as Anowar and Sadaoui (2020) and Batageri and Kumar (2021).

2. Credit Card Fraud datasets tend to be heavily imbalanced. There are differences in the literature on how to take concrete steps to tackle this problem and avoid model bias. Priscilla and Prabha (2020) propose that resampling techniques themselves could be distorting credit card fraud data, which will impact on downstream results, including XAI outputs.

Gaps: How exactly does a researcher measure and display 'explainability' in Explainable Artificial Intelligence Research?

1. In their research experiments with the LIME (**L**ocal **I**nterpretable **M**odel-agnostic **E**xplanations) algorithm, Ribeiro et al. (2016) describe how users can have a *trust* issue with ML models, like NN, that 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), Lundberg et al (2017), Honegger (2018), and Sinanc et al. (2021). There appears to be no cast iron process to ensure this trustworthiness.

2. 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 (2021) 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, there is no well established output framework for explaining credit card fraud classification through 'black-box' models.

3. The 'If-Then' style of rules could be an alternate XAI output option to be chosen for this dissertation. Vilone & Longo (2021) also assert that there is still relatively little research that objectively assesses this approach with quantitative metrics, thus allowing it to be benchmarked against other XAI methods.

4. Psychoula et al (2021) state that the runtime implications of XAI output (explanations) 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.

5. 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.

Research Question: To what extent can we quantify the quality of contemporary machine learning interpretability techniques in the classification of credit card fraud transactions by a 'black box' Neural Network ML model?

Hypothesis + Research Methods

Null Hypothesis

A conventional view is that the workings of credit card fraud detection Neural Network models are a 'black-box' process, and it is difficult to quantify the best interpretation framework to explain the reason for a given classification result.

Alternate Hypothesis

IF I train a Neural Network algorithm for ML credit card fraud detection, and apply different interpretability frameworks to the model results
THEN then I can measure the output of each framework against a set of metrics (slide 8), acting as unified quantitative measure, and determine the statistically best approach to explaining local, post-hoc credit card fraud classification results.

Research Methods

This will be a ***primary research*** approach, based on insights from a review of certain literature in the field of XAI research.

The ***objective*** is to conduct a sequence of lab experiments to measure the empirical performance of different interpretability frameworks on a NN model built for credit card fraud detection.

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 will be a ***deductive*** approach to test the assumption that one particular interpretability frameworks can be shown, through the numerical outputs of each experiment, to generate the best local explanations for a credit card fraud classification result.

General + Specific Research Objectives for experimental purposes towards hypothesis testing using statistical tools

Research Aim:

- To rank selected interpretability frameworks (LIME, SHAP, LORE, Anchors, and InterpretML), using predefined metrics, against the output from a NN credit card fraud detection model and determine which one, if any, demonstrates the best overall performance.

General / Specific Research Objectives

- **01:** *Pre-process credit card fraud dataset to improve interpretability measurement. (Internal company dataset has already been provided).*
  - 15% of records in dissertation dataset are labelled 'fraud'. Produce a 50/50 balanced training and test dataset by removing appropriate number of 'non-fraud' records.
  - Reduce dimensionality of data (Ribeiro et al., 2016). Remove highly correlated features and limit to top 20 features based on a feature importance ranking by an RF algorithm. Generate a new dataset for experimentation.
- **02:** *Train and test NN model for credit card fraud detection.*
  - Partition data set into 80% training / 20% testing.
  - Use ANN algorithm to generate model on training data. Validate F1 and Recall scores produced by model against the test data. Refine model parameters if necessary to achieve expected model performance criteria (slide 8),
- **03:** *Produce explanations for model predictions with each framework.*
  - In separate experiments, use LIME, SHAP, LORE, Anchors, and InterpretML to generate explanations for model predictions for each instance in the test data.
- **04:** *Differentiate the performance of each interpretability framework. (ElShawi et al., 2020)*
  - Use pre-defined metrics (slide 8) to grade each framework. Determine if one framework demonstrates a clear numerical superiority across all metrics.
- **05:** *Summarise learnings from experiments to compare interpretability frameworks.*
  - Explain rationale for conclusions to research. Propose areas of further study.

Bibliography

Adadi, A., & Berrada, M. (2018). Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, *6*(52), 138–160. https://doi.org/10.1109/access.2018.2870052

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. https://doi.org/10.1109/smc42975.2020.9283136

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

Dal Pozzolo, A., Caelen, O., Le Borgne, Y.-A., Waterschoot, S., & Bontempi, G. (2014). Learned lessons in credit card fraud detection from a practitioner perspective. *Expert Systems with Applications*, *41*(10), 4915–4928. https://doi.org/10.1016/j.eswa.2014.02.026

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

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

Honegger, M. (2018, August 15). *Shedding light on Black Box Machine Learning Algorithms: Development of an axiomatic framework to assess the quality of methods that explain individual predictions*. arXiv.org. Retrieved December 4, 2022, from https://arxiv.org/abs/1808.05054v1

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). essay, NeurIPS Proceedings.

Priscilla, C. V., & Prabha, D. P. (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. https://doi.org/10.1109/icssit48917.2020.9214206

Bibliography

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. https://doi.org/10.1109/mc.2021.3081249

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "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. https://doi.org/10.1145/2939672.2939778

Sharma, A., & Bathla, N. (2020). *Review on Credit Card Fraud Detection and Classification by Machine Learning and Data Mining Approaches*, *6*(4), 687–692. Retrieved from https://www.semanticscholar.org/paper/Review-on-credit-card-fraud-detection-and-by-and-Sharma-Bathla/b6c839cadb4c6281a934a8788fec93d5482e6af4.

Sharma, P., & Priyanka, S. (2020). 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. https://doi.org/10.35940/ijeat.e9934.069520

Sinanc, D., Demirezen, U., & Sağıroğlu, Ş. (2021). Explainable Credit Card Fraud Detection with Image Conversion. *ADCAIJ: Advances in Distributed Computing and Artificial Intelligence Journal*, *10*(1), 63–76. https://doi.org/10.14201/adcaij20211016376 A new explainable artificial intelligence approach is ... presented. In this way, feature relationships that have a dominant effect on fraud detection are revealed.

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

Vilone, G., & Longo, L. (2021). Classification of explainable artificial intelligence methods through their output formats. *Machine Learning and Knowledge Extraction*, *3*(3), 615–661. https://doi.org/10.3390/make3030032

Vilone, G., & Longo, L. (2021). Notions of explainability and evaluation approaches for Explainable Artificial Intelligence. *Information Fusion*, *76*, 89–106. https://doi.org/10.1016/j.inffus.2021.05.009

Performance Metrics for Experiments

Explainability Metrics *(based on explainability framework comparison research by (Guidotti et al., 2019); (Honegger, 2018); (ElShawi et al., 2020)*;

1. *Fidelity.* A measure of the matching decisions from the interpretable predictor against the decisions from the 'black box' model.
2. *Stability*. Instances belonging to the same class have comparable explanations. K-means clustering 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*: 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. *Time*: Average time taken, in seconds, by the interpretability framework to output a set of explanations. (Similar Cloud environments are applied to all experiments).

Metrics to apply to any meaningful credit card fraud detection model;

1. ***F1*** and ***Recall*** are better score for credit card fraud detection problems, as opposed to simple accuracy, because of the uneven class distribution seen in many credit card datasets. Taking comparative NN fraud detection experiments from Sinac et al. (2021) and Anowar & Sadaoui (2020), a target threshold of **>= 0.85 and >=0.9** will apply for F1 and Recall, respectively, to the NN model created in the initial experiment steps. This will ensure that a performant NN model has been created prior to the measurements of the results from the five experiments on the separate interpretability frameworks.

Fw: Assignment 2 - research design and proposal writing - feedback

D21124026 Ciaran Finnegan

To: ciaran@feefinnegan.com; ciaran.finnegan@netreveal.ai

From: Luca Longo <Luca.Longo@TUDublin.ie>
Sent: Sunday, December 11, 2022 3:40 PM
To: D21124026 Ciaran Finnegan <D21124026@mytudublin.ie>
Subject: Assignment 2 - research design and proposal writing - feedback

domain (2.5%)	issues in (Dom=domain, S=scope, L=limitations, D=delimitations)	gaps (2%)	RQ (2%)	hypothesis (3.5%)	research methods (1%)	research objectives (5%)	apa (2%)	dataset (2%)	mark 2nd assignment
1	S,A,L	2	2	2	1	3	2	2	15

Dr. Luca Longo
Artificial Intelligence & Cognitive Load Research Lab (AICL-lab)