1. **Front-page**  
   * Student name and surname
   * Student number
   * Stream (eg. ASD, DA) or PhD structured program
   * DIT programme (eg. DT228A / DT228B, )
   * Title of proposal (max words 20 words)
   * List of people and companies associated with the project (including those you might need input from or data from) (max 20 words)
   * Sources of data needed for the project, if any (max 30 words)

Title:

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

Dataset:

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

1. **The research background**(max 300 words).  
   Describe the context of application and provide the reader with some background and main notions/concepts related to the domain and scope of your research.

Mention the usual ‘black box’ concerns – contemporary models need to be performant but may sacrifice transparency as a result. LOOK For REFERENCES?

Reference GDPR – typifies the focus on auditing, internal and external. LOOK For REFERENCES?

Focus also on need to objectively assess different explainer techniques, as part of product roadmap considerations for commercial fraud detection products.

Should scope broaden beyond NN – reference EBM in MicrosoftML.

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.

1. **Informal description of the research problem**(max 300 words)  
   Describe the problem you aim to tackle informally. Also define the assumptions, limitations and delimitations of your research proposal.

Informal description of problem to be tackled: Objective assessment of state-of-the-art ML explainers, as applied to credit card fraud detection. Compare a set of common XAI techniques and look for insights into the relative strengths of each technique.

Is there a viable ‘glass-box’ alternative DNN models for credit card fraud detection?

Focus also on need to objectively assess different explainer techniques, as part of product roadmap considerations for commercial fraud detection products.

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

1. **Literature review and its gaps + state-of-the art approaches to solve the identified research problem**(max 1000 words)  
   Describe the relevant peer-reviewed articles you have read (21+) in the selected domain of research and identify the gaps. Additionally, identify and describe the state-of-the-art approaches to solve the identified research problem.  
   Gaps and state-of-the-art will inform the proposal of your 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 ste ps 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 (Local Interpretable Model-agnostic Explanations) 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 introducti on 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.

Describe State-of-the-art approaches (take from paper 31 + 58)

SHAP

LIME

ANCHOR

InterpretML (EBM)

1. **Research question**(max 70 words)

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, but also considers this output against the context of an additional ‘glass-box’ explainer.

1. **Hypothesis**(max 300 words)  
   Formally define your alternate and null hypothesis as well as provide a textual description of your alternative hypothesis.

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.

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 EBM.

Alternate Hypothesis:

IF a Neural Network algorithm is trained for ML credit card fraud detection in parallel with the creation of a ‘glass-box’ EBM model, and SHAP, LIME, ANCHORS, and EBM 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 pre-defined set of metrics, to rank each explainer technique and demonstrate statistically which is best for explaining local credit card fraud classification results.

Section 7 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 EBM. 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.

Establishing a ranking 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, particularly as it an objective to validate the assumption from Microsoft researchers that their EBM technique is as accurate as black box models. (‘Glass box’ models are often perceived as less performance than NN, so this will be a key supplementary observation of these experiments). A Wilcoxon signed-rank test will be applied pairwise on the interpretability techniques to measure the scale of difference, if any, in performance.

1. **Research objectives and experimental activities**(max 1000 words)  
   Define your general objectives and their specific research objectives. (you can use a combination of textual and visuals). Provide precise details about how you plan to achieve each specific research objective (eg. programming languages, technologies employed, execution of surveys, baseline methods/approaches etc.).In some of the specific research objectives, you might need to clearly specify all the details of dataset (eg. dependent/independent variables, scales and ranges, sample size etc.) as well as specific formulas. The overall goal is to allow your reader to implement your experimental design and independently achieve each research objectives.

Title:

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

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.

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

• O1: Pre-process credit card fraud dataset to improve interpretability measurement. (Internal company dataset has already been provided).

o 15% of records in dissertation dataset are labelled ‘fraud’. Produce a 50/50 balanced training and test dataset by removing appropriate number of ‘nonfraud’ records.

o 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.

• O2: Train and test NN model for credit card fraud detection.

o Partition data set into 80% training / 20% testing.

o 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),

• O3: Produce explanations for model predictions with each framework.

o In separate experiments, use LIME, SHAP, LORE, Anchors, and InterpretML to generate explanations for model predictions for each instance in the test data.

• O4: Differentiate the performance of each interpretability framework. (ElShawi et al., 2020)

o Use pre-defined metrics (slide 8) to grade each framework. Determine if one framework demonstrates a clear numerical superiorit y across all metrics.

• O5: Summarise learnings from experiments to compare interpretability frameworks.

o Explain rationale for conclusions to research. Propose areas of further study.

1. **Evaluation of designed solution with performance metrics (and statistical tests)** (max 300 words)  
   Describe carefully how you are going to evaluate the outcomes of your experiment statistically, the performance metrics you have planned to use, considering the concept of significance, and how you are going to accept/reject your hypothesis  
   Describe how findings will be related to the research question

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.

Add description of Significance (Friedman) tests, and Reference…