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:

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 on a credit card transaction dataset, in parallel with the creation of a ‘glass-box’ EBM model, for ML fraud detection, 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 similarity 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.

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, particularly as it is an objective to validate the assumption from Microsoft researchers that their EBM technique is as accurate 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.

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 superiority across all metrics.

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

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

The study will be carried out in three parts. A very large case base is required so that results can be collected for many different sized case bases. A paper investigating CRNs found performance issues for case bases of 40,000 cases, a data set of at least 1 million will be used so that the case-base size can be gradually increased to show how the performance degrades. The content of the case base is not that important, the only requirement is that it is large enough. The first stage will use the k-Nearest Neighbour classifier on the case base to show the existence of the utility problem. The next stage will apply the CRN method to the same case base to show some improvement on the linear approach but will still exhibit the utility problem for larger case-bases. The final stage will apply the big data method to the case base. The big data platform that will be used will be Hadoop, and it will use the programming model MapReduce to execute the CBR retrieval step. The first step will use a very small sample of the dataset, one which will not suffer from the utility problem. The size of the case base will be increased for each new retrieval step and the performance measured as the time taken to retrieve the cases from the case base. The utility problem will occur as the case base gets larger and performance will deteriorate. The time taken to retrieve cases from the case base will be recorded to measure performance and the degradation in performance can be described with a graph which shows time taken on the x-axis and number of cases on the y-axis. CRNs will be used on the same sample sizes with the same query so that the performance can be measured on retrieval time. There will be an improvement in performance, as has been shown in previous literature, but as the dataset becomes very large the utility problem will still occur. The data will be plotted on the same graph as the kNN results to show the difference in speeds between the two methods. MapReduce is a programming model which can process large datasets. A map function is specified which processes key/value pairs to generate a set of intermediate key/value pairs and so would be an ideal solution for a case-based reasoning problem. Hadoop MapReduce can process very large amounts of data in parallel on large clusters. The same query will be carried out using this method on the same sample sizes and the time take to retrieve the cases recorded. The performance of the three methods can then be compared by plotting the results of this stage of the experiment alongside the previous two graphs. The UCI repository of machine learning has several data sets that are larger than 1 million which would be suitable for the proposed evaluation. One such data set is the Knowledge Discovery and Data Mining Tools Competition Data for 1999. This dataset contains simulated intrusions in a military network environment and the problem to solve would be whether a new connection would be considered “good” or “bad” based on the cases in the database. The data set has 4 million entries, which will be sufficient for the problem proposed.

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

The study will execute five primary research steps to build up a table of metrics for each explainer method and allow a statistical comparative analysis of the performance by each explainer technique. The research focus is on explanations for fraud classification of individual transaction records – hence these experiments only consider *local, post-hoc* results. A *global* analysis of how the model classifies fraud is not in scope.

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. Much of this data has a very tenuous impact on the fraud classification process, and dimensionality reduction and feature engineering will be key initial steps in the design of the initial experiment steps in this paper.

However, 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.

For the creation of the credit card fraud detection model 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 my detail in Section 8 of this submission document);

1. Fidelity.

2. Stability.

3. Separability.

4. Similarity.

5. Time

<Image> Adapt Luca’s paper…

A very peripheral objective of this research is to assess the ease of use of cloud-based ML development options for this type of work. 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 ‘Time’ metric can be compared accurately across all explainer techniques.

As described above, 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 dimensionality process, and it will be 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 first model building exercise will be written in Python and 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 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 this new NN model. 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 second model exercise begins with another SageMaker notebook importing the Microsoft InterpretML ‘interpret.glassbox’ libraries and building a fraud classifier with the Explainable Boosting Machine (EBM) algorithms. Similar performance metrics will be expected on the test set, as referenced above for the ANN model.

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, and ANCHORS explainability techniques are used to generate the explanations from the ANN model. The InterpretML library is used to generate EBM explanations.

A table of metric scores for each explainability technique will be compiled for each batch of predictions. This approach follows some of the concepts in measuring similarity performance for explainability techniques as elaborated by ElShawi et al (2020). A significance test will be run on the tabular data to determine if certain approaches are better at explaining credit card fraud. The statistical analysis will

• 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 superiority across all metrics.

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

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

Experiment Step 1…

Experiment Step 2…

Experiment Step 3…

Experiment Step 4…

Experiment Step 5…

Conclusion…

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

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.

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).

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

Relate findings to research question…