# 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 (SHapley Additive exPlanations) 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.