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

- O1: *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.
- O2: *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),
- O3: *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.
- O4: *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.
- O5: *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.