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 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 InterpretML (EBM) 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, and ANCHORS interpretability methods upon a Neural Network model trained on a commercial dataset containing credit card transactions, which are labelled *'fraud'* or *'non-fraud'*. The analysis is then continued with the inclusion of the EBM algorithm from Microsoft's InterpretML library. This step is done to ask the question; is there a viable *'glass-box'* alternative to ANN models for credit card fraud explanations? 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 is 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'* 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 some of the most common 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 step 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 also 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 InterpretML (EBM)

Microsoft InterpretML is an open-source Python package containing libraries of machine learning interpretability algorithms (Kaur et al., 2020). It exposes both black box and *'glass box'* interpretability techniques under a unified API. Of interest to this dissertation is a new glass box interpretability model called Explainable Boosting Machine (EBM).

The core of EBM is a round-robin procedure to train on one feature at a time using a low learning rate and show how each feature contributes to the model's prediction for an instance.

Microsoft claim that EBM can deliver an accuracy and performance comparable with black box models such as ANN, while also delivering highly intelligible explanations. It has been included in the experiments proposed in this dissertation to determine if EBM scores can objectively compete with the SHAP, LIME, and ANCHORS explanations on a NN model for credit card fraud predictions.

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

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 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 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 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.

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 InterpretML), using predefined similarity metrics, against the output from Neural Network (NN) and Explainable Boosting Machine (EBM) credit card fraud detection models 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 8 of this submission);

1. Fidelity
2. Stability
3. Separability
4. Similarity
5. Time

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 'Time' 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 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

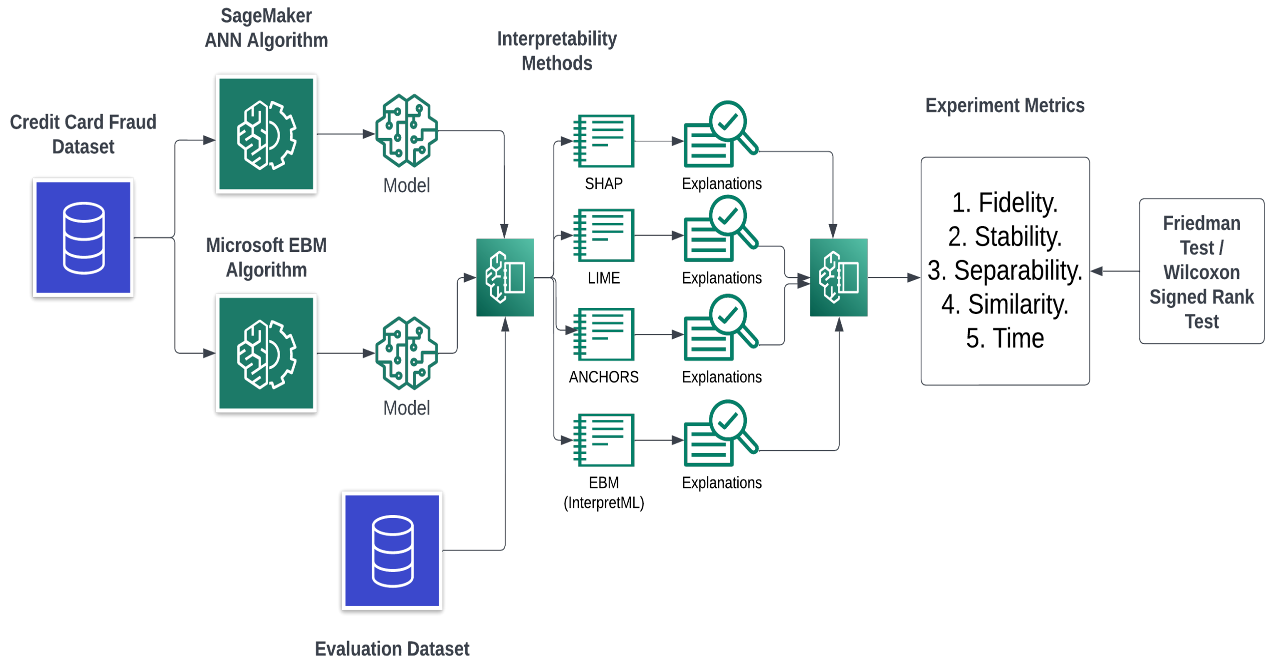


Figure 1: *Overview of experiment design*

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 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 Sinanc et al. (2021) and Anowar & Sadaoui (2020), a target threshold of **¿= 0.85** and interpret.glassbox will apply for **F1** and interpret.glassbox, 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 EBM algorithm. 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.

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 explainability 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. 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).

A Friedman test will be run to determine if evidence exists that there is a difference in performance between SHAP, LIME, Anchors, and EBM 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.

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.

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

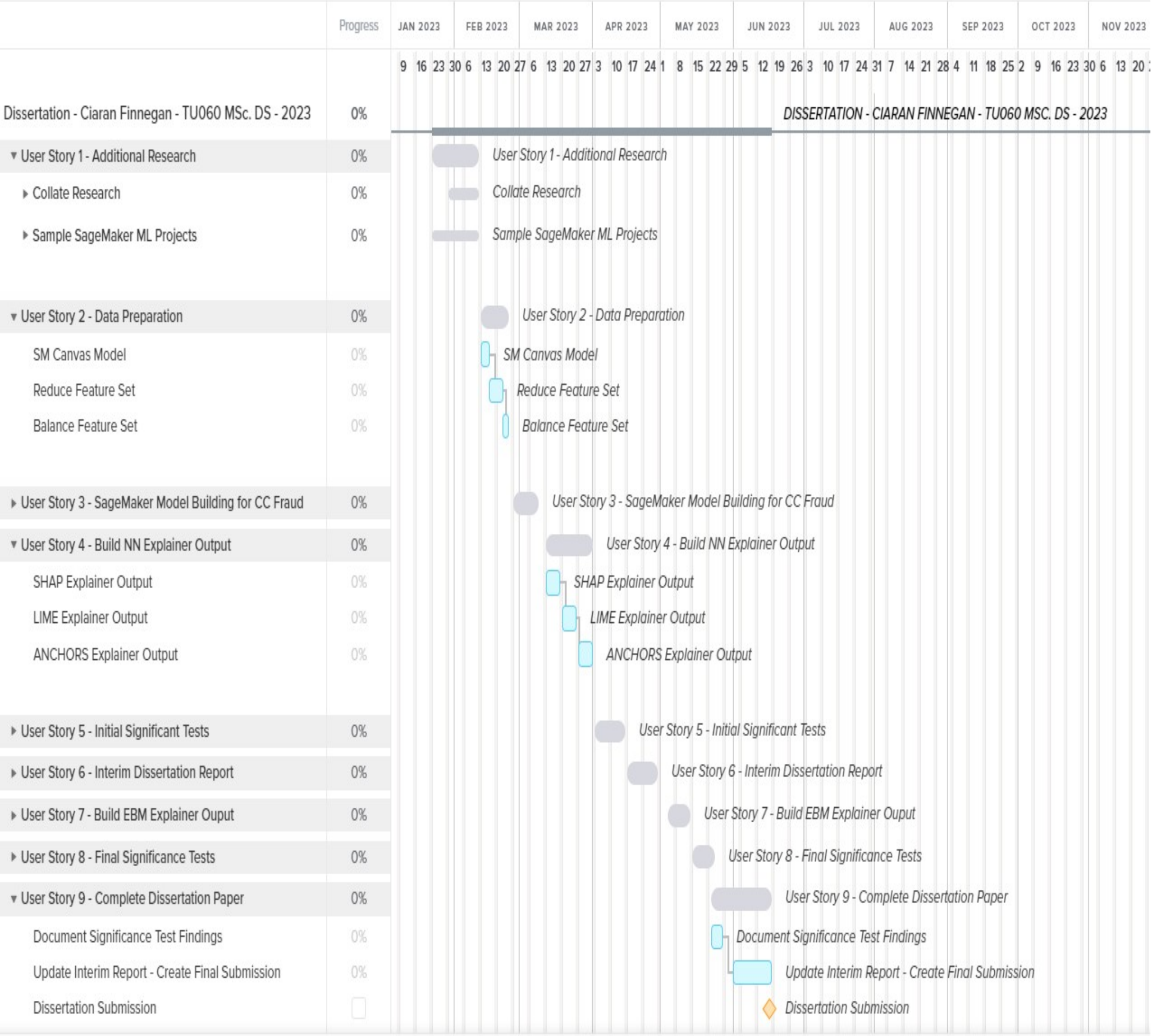
Week 4 is the model building activity (User Story 3). One model will be built with a SageMaker ANN algorithm, the other will be created using libraries imported from the InterpretML GitHub repository. User Story 4 will focus on the most complex and lengthy period of dissertation activity; producing the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

User Story 5 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.

The next experiment introduces the 'glass box' EBM model and 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 shows the Gannt Chart with timescales for the research activities in this dissertation.



	Progress
Dissertation - Ciaran Finnegan - TU060 MSc. DS - 2023	0%
▾ User Story 1 - Additional Research	0%
▸ Collate Research	0%
▸ Sample SageMaker ML Projects	0%
▾ User Story 2 - Data Preparation	0%
SM Canvas Model	0%
Reduce Feature Set	0%
Balance Feature Set	0%
▸ User Story 3 - SageMaker Model Building for CC Fraud	0%
▾ User Story 4 - Build NN Explainer Output	0%
SHAP Explainer Output	0%
LIME Explainer Output	0%
ANCHORS Explainer Output	0%
▸ User Story 5 - Initial Significant Tests	0%
▸ User Story 6 - Interim Dissertation Report	0%
▸ User Story 7 - Build EBM Explainer Ouput	0%
▸ User Story 8 - Final Significance Tests	0%
▾ User Story 9 - Complete Dissertation Paper	0%
Document Significance Test Findings	0%
Update Interim Report - Create Final Submission	0%
Dissertation Submission	□

Figure 2: *Gannt Chart*

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

Kaur, et al. (2020, Apr). Interpreting interpretability: Understanding data scientists' use of interpretability tools for machine learning. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–14. doi: 10.1145/3313831.3376219

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