Literature Review for Linear and Non-Linear Models Final Project

Course Instructor: Dr. Utku Pamuksuz

Prepared by:

Dharti Seagraves

Steve Veldman

Prashant Kulkarni

ADSP 31010 ON02

February 2024

Table of contents

Introduction

This document is prepared to support the project of Bank transactions fraud detection. We reviewed various research papers from IEEE and Google Scholar and compiled this document to support our choices in final project implementation. This document discusses three research papers in which various methods used by scholars and researchers showcase the effectiveness of fraud detection techniques using machine learning algorithms. Our efforts have been to understand the choices of algorithms available to us, various tests to measure the effectiveness of those choices and to use them in our final project.

The research goes beyond the algorithms taught in this course and may include others typically used in earlier or later phases of a machine learning project. We also explored the methods used for Explainable AI (XAI) considering its importance to consumers and requirements of regulations in BFSI. We provided a comprehensive review of these methods , their suitability and a layered approach of XAI.

We have also included review of techniques used for handling imbalanced data like Synthetic Minority Over-sampling Technique (SMOTE) in support of our project.

Research paper 1

Article: A machine learning based credit card fraud detection using the GA algorithm for feature selection“[1]

Published in: A machine learning based credit card fraud detection using the GA algorithm for feature selection | Journal of Big Data

The paper emphasizes the significance of feature selection in ML-based fraud detection and introduces the use of **Genetic Algorithms (GA)** to improve this process. We selected this paper because it provides a review of various algorithms used in the fraud detection place.

GA stands for Genetic Algorithm. It is a search heuristic that mimics the process of natural selection. The study employs various classifiers including Decision Trees, Random Forest, Logistic Regression, Artificial Neural Networks, and Naive Bayes.

Feature selection (FS) is identified as a critical step in the implementation of machine learning methods. The significance of FS stems from the potential large feature space present in datasets used during the training and testing processes, which could negatively impact the overall performance of models.

Various citations include GA based FS to improve the performance of intrusion detection systems (Kasongo[2]), Stacked sparse autoencoder network (SSAE) (Mienye et al.[3]), Enhanced Principal component analysis (Hemavathi et al.[4]) and hybrid FS (Pouramirarsalani et al.[5]). Random Forest (RF) is used as the fitness method within the GA, chosen for its ability to address overfitting issues common in Decision Trees (DTs), handle both continuous and categorical attributes, and perform optimally on datasets with class imbalance.

An instance of the RF classifier is instantiated and trained using the training set on a normalized input. The resulting model is evaluated using the testing data, and predictions are stored. The evaluation process assesses the model's performance based on the predictions. The fitness measure is determined by the accuracy a particular attribute vector achieves in the testing process.

Metrics used to evaluate the models' performance, such as accuracy, precision, recall, and F1 score. The Authors discuss the results of experiments, highlighting how the GA-based feature selection improved the performance of classifiers, with Random Forest achieving the best results.

Key findings of the paper includes:

- The GA-Random Forest model, using the v5 feature vector, achieved the highest overall accuracy of 99.98%. Other classifiers also showed remarkable accuracy, with the GA-DT model achieving 99.92% accuracy using the v1 feature vector.
- The experiments demonstrated superior results compared to existing methods, with the GA-RF model outperforming others by significant margins.
- The paper also validates the proposed framework on a synthetic credit card fraud dataset, showing that the GA-DT model achieved an AUC of 1 and 100% accuracy, followed by the GA-ANN model with an AUC of 0.94 and 100% accuracy.

The study concludes that the GA-based feature selection method significantly enhances the performance of machine learning classifiers in detecting credit card fraud.

Research paper 2

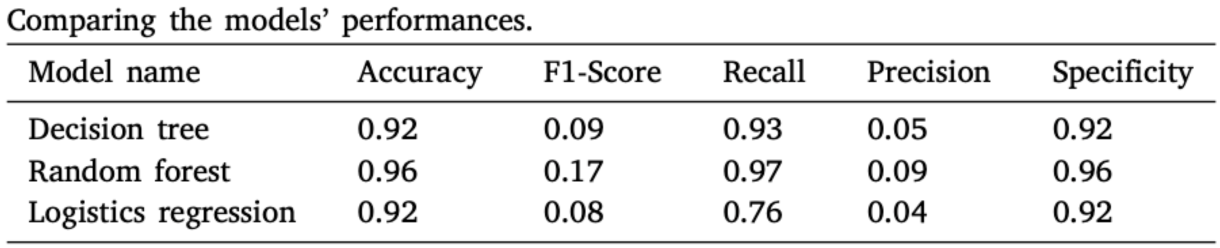
Article: A supervised machine learning algorithm for detecting and predicting fraud in credit card transactions“[3]

Published in:
https://www.sciencedirect.com/science/article/pii/S2772662223000036#b29

This article provides a machine learning approach to detect and predict fraud in credit card transactions. The authors studied logistic regression, decision trees, and random forests to classify, predict, and detect fraudulent transactions using simulated data from January 1, 2020, to December 31, 2020.

With an increase in online transactions, fraud targeting has also risen, particularly affecting individuals over 60 during the late-night hours of 10 pm to 5 am. The use of VPNs by fraudsters complicates the timely identification of fraudulent activity, making it crucial to detect fraud at the transaction's occurrence.

The authors used various models for fraud detection, highlighting the Gradient Boosting method for feature importance analysis, achieving 95.9% accuracy. After preprocessing, which included handling missing values and balancing the dataset with SMOTE, the study compared the effectiveness of Logistic Regression, Decision Trees, and Random Forest models. With 92% accuracy for logistic regression and decision trees and 96% for random forests—which also had the highest AUC of 98.9%—the study determined Random Forest as the best model for predicting fraudulent transactions based on various performance metrics.



Comparing the models' performances.

Model name	Accuracy	F1-Score	Recall	Precision	Specificity
Decision tree	0.92	0.09	0.93	0.05	0.92
Random forest	0.96	0.17	0.97	0.09	0.96
Logistics regression	0.92	0.08	0.76	0.04	0.92

Additionally, the data revealed that most fraudulent transactions occurred on Sundays, late at night when potential victims were likely asleep, and banks were closed. The authors suggest providing more in-person services to older clients and enhancing security measures during these vulnerable hours. They recommend financial institutions implement the Random Forest model for fraud detection, a strategy that could also extend to healthcare and other industries facing classification and prediction challenges.

Research paper 3

Article: Credit Card Fraud Detection using Artificial Neural Network and Back Propagation

Published In: 2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS)

This article addresses the topic of credit card fraud detection using artificial neural networks (ANN). It begins with a literature review highlighting previous research evaluating ANN against logistic regression, the use of cost-sensitive modeling to address the issue of data imbalance, comparison of multilayer perceptron neural networks with Naïve Bayes and Decision Tree models, and ANN utilizing Self-Organizing Maps (SOM). The authors note that out of these previous studies, models utilizing SOM performed the best but would likely have significant latency issues when scaled to the level of the banking system.

The authors propose an ANN using gradient descent and backpropagation, with the aim to perform accurate classification in real-time and at scale. They utilized the Credit Card Customer dataset available on Kaggle, which contains 30 predictor variables. For confidentiality the dataset has already been transformed using PCA, and the resulting principal components are what was used to train and evaluate this model. The authors segregated the original data set into 80% for training the model, 10% for testing the model, and 10% for final validation. Their model uses an input layer, three hidden layers with 15 neurons each (calculated based on the number of neurons in the input layer), and an output layer. The RELU activation function was used for the hidden layers, and the sigmoid activation function was used for the output layer.

Compared to several other articles that I read on the use of neural networks for fraud detection, this article stood out for its level of detail in explanation both the underlying concepts used and the design of this specific model. The article provides a thorough explanation of how the proposed ANN functions, as well as detailed documentation of the underlying formulas. The process of how gradient descent and backpropagation are used to work through the hidden layers of the neural network in reverse order is also explained very clearly. The authors not only provide more detail on the structure of their proposed model, they also provide a very useful 7-step description of how the model was trained.

This model achieved an admirable 99.92% accuracy, 99.96% recall, and 99.96% precision for an F1 score of 99.96%. While this is impressive model performance, these results do not address one important shortcoming – the model failed to identify 11 out of the 55 fraudulent transactions in the validation set, which means that in production ~20% of credit card fraud would potentially pass undetected. This highlights a widespread shortcoming that I noticed across the literature on this subject. When evaluating these models, a great deal of emphasis is placed on traditional metrics without much discussion of the practical implications for this specific use case.

In this specific instance, it Is likely that the large number of undetected fraudulent transactions is related to the fact that data set used was incredibly imbalanced (they do not provide an exact percentage for the overall data set, but in the validation set only 0.193% of observations were from the positive class). It does not appear that the authors used any techniques to address this issue, which I find to be the major shortcoming of this article. Even if the authors had determined that this was out of scope for this particular effort, I would have liked to have seen this mentioned as a line of inquiry for future work.

Explainable AI methods

Article: Explainable AI: current status and future directions
Published in : IEEE https://arxiv.org/pdf/2107.07045.pdf

The paper provides an extensive review of various explainable AI (XAI) techniques proposed for different data types like image, text, audio and video. It categorizes these techniques into - transparent methods, post-hoc methods, model-agnostic and model-specific methods.

The suitability and applicability of these methods depends on factors like model complexity, data dimensions, feature dependencies and end-user requirements.

Transparent Methods: These include inherently interpretable models like linear regression, decision trees, Bayesian networks. They are best suited for applications with simpler data relationships and limited feature interactions. Key strengths are algorithmic transparency and human-simulatability. But their simplicity can compromise accuracy.

Post-hoc Methods: Techniques like LIME, SHAP, LRP are flexible to apply on complex black box models like neural networks. They approximate model behavior by sampling data points or back-propagating relevance scores. Post-hoc methods enable model-agnostic and local explanations but can be computationally expensive.

Model-Agnostic: LIME and SHAP are model-agnostic, applicable across model types. They provide local feature importance estimates via data sampling and perturbation. Limitations are instability and neglect of feature interactions. Model-specific methods can better incorporate model and data intricacies.

Model-Specific: Methods like Layer-wise Relevance Propagation (LRP), counterfactual search leverage model internals for tailored explanations e.g highlighting influential neurons/layers. But they require access to model internals.

Let us review the specific XAI method and their limitations:

- LIME (Local Interpretable Model-Agnostic Explanations) - LIME is a model-agnostic method that approximates a complex model locally with an interpretable model to explain individual predictions. It is flexible across data types. However, explanations are local and linear models may not fully capture complex relationships.
- SHAP (SHapley Additive Explanations) - Based on game theory, SHAP assigns each feature an importance value for a particular prediction. It accounts for feature interactions unlike simpler approaches. But it can be computationally intensive for some models.
- Counterfactual Explanations - Counterfactuals explain model outputs by showing how changes to the input would alter the output. This technique is intuitive but

generating useful counterfactuals is challenging, particularly for high-dimensional and discrete data.

- Explainable Boosting Machines (EBM) - EBMs combine interpretable decision trees with advanced boosting algorithms for state-of-the-art accuracy while retaining explainability. However, individual trees can still be complex and difficult to parse.
- General Additive Models (GAMs) - GAMs model relationships between inputs and outputs via interpretable smoothing functions rather than raw feature importance scores. GAMs make simplifying assumptions, struggling with some complex data.

Here are key tradeoffs that should be considered between different categories of XAI methods:

Accuracy vs Interpretability:

There is often an intrinsic tradeoff between the accuracy of a model and how interpretable it is. Highly complex models like large neural networks tend to have superior predictive accuracy but their internal logic is not intuitive or transparent. On the other hand, simpler transparent models are easier to understand but typically underperform on benchmark metrics.

The choice depends on the acceptable level of accuracy degradation for the use case - safety-critical applications may warrant some loss of performance for trust and accountability. Post-hoc methods aim to balance both but approximating complex model behavior has limitations regarding stability and fidelity.

Global vs Local Explanations:

Global interpretability methods provide an understanding of the complete model behavior for the entire input space. But they can become exponentially complex for high-dimensional data. Local explanation techniques highlight influential features for individual predictions, keeping explanations focused and granular. However, they may miss global interactions and lack coverage guarantees.

Model-specific vs Model-agnostic:

Model-specific methods can deeply probe the internals of particular model architectures and data pipelines, providing meticulous and tailored explanations. However, these are not generalizable across model types. Model-agnostic techniques treat the original model as a black box but allow flexibility across model selection. The fidelity vs flexibility spectrum needs to be considered.

Static vs Interactive Explanations:

Some methods generate a one-time static explanation for predictions. Interactive interfaces allow changing feature values and re-generating explanations on the fly for what-if analysis - more intuitive for users but computationally intensive. The right balance depends on latency constraints and user requirements.

Overall, XAI method choice entails navigating accuracy, generality and resource overhead tradeoffs while aligning to use case needs regarding trust, causality, interactivity etc. Trade-offs exist between accuracy, flexibility and level of explanations across these categories. No single method addresses all XAI challenges. A layered approach is often helpful. Let us take a look at this.

Let us consider a computer vision classification model. The internal operations are not transparent, so a post-hoc method like SHAP is used to reveal influential pixels and visual features driving predictions. While intuitive, raw pixel importance scores have limitations. So counterfactual image perturbations are also generated to validate causality and check sensitivity.

On top of this, if the users are clinicians, an interactive interface allows tweaking feature attributions and pixel heatmaps in real time to build appropriate mental models regarding model behavior. And finally, complementary techniques add textual descriptions of imaging biomarkers in a format domain experts comprehend.

This **layered approach** tackles various challenges in realizing trustworthy and actionable XAI systems:

- Provides multimodal explanations aligning both technical users and domain experts
- Checks fidelity and stability using alternate feature importance estimations
- Improves mental model development via interactivity
- Handles accuracy-interpretability and local-global tradeoffs
- Allows flexibility across models and tasks through model-agnostic techniques

Thus, combining XAI methods based on their individual strengths and weaknesses extracts complementary benefits. The big picture helps choose techniques tailored to unique use case goals.

Implementation Considerations:

That said, effectively implementing layered XAI also introduces engineering challenges - explanations must integrate cleanly, resources need to scale, and accuracy should not degrade substantially. A principled framework is required for combining methods lacking native interoperability. But overall, this direction shows promise in operationalizing impactful XAI.

Here is one example of how principled frameworks like Fusion Engine can be used:

Fusion Engine is a specialized middleware layer that consumes explanations generated from heterogeneous XAI components and consolidates them into an integrated narrative presented to end users or domain experts. It handles low-level aggregation while resolving conflicts.

For example, let's say we have an image classification pipeline with 3 components. Each generates self-contained explanations with little native interoperability:

1. CNN Model: Predicts image categories along with heatmaps highlighting predictive regions using Gradient-SHAP method
2. Counterfactual Generator: Produces minimally edited image versions that alter CNN predictions.
3. Ontology Mapper: Annotates influential pixels and attributes domain concepts for user context.

The fusion engine ingests them via standardized schemas and performs aligned fusion into an integrated visual overlay tagged with textual callouts. It merges heatmaps, resolves annotation differences using ontology linkage and evaluates stability using counterfactuals - finally presenting a consolidated explanation.

The fusion engine can be built using:

1. End User Facing Interface: This provides the consolidated explanation to users by aligning different modalities like visuals, texts, data visualizations etc.
2. Fusion Engine: The core middleware logic that ingests heterogenous explanations, evaluates them, resolves conflicts and aggregates them into a unified narrative.
3. XAI Modules: The different explanatory components like CNN heatmaps, counterfactual generators, ontology mappers that analyze the ML model and data. These are packaged as independent containerized microservices.
4. Orchestration Layer: Leverages platforms like Kubernetes to manage and scale containers as well as handle deployment.
5. Interoperability Layer: Relies on API schemas and protocols for clean interplay between the explainer modules and fusion engine.

Thus a structured fusion approach enables realizing complimentary benefits of different interpretable capabilities in a scalable and reusable manner.

Now that we have gone through a overview of XAI methods, trade offs, here are various use cases XAI can be applied:

- **Computer Vision:** Post-hoc attribution via LIME and SHAP help highlight relevant pixels and image regions that influence classification decisions e.g. salient edges and textures. Counterfactual search finds minimal evidence for target class predictions.
- **NLP:** Feature/concept importance scores bring interpretability to text classification models. Attention layers also indicate influential tokens. Declarative rule/tree representations add transparency for end users.

- **Time-series:** Judicious perturbations on sensor signals passing through complex forecasting models can reveal feature importance for predictions. Domain knowledge integration can aid human simulatability.
- **Recommendation:** Techniques like SHAP add transparency by identifying key user attributes and item features driving recommendations. This builds user trust and aids debugging.
- **Healthcare:** Method combinations provide multimodal explanations - saliency maps reveal predictive ROIs in scans, rules/trees describe clinical guidelines encoded by models. Improves physician trust and accountability.

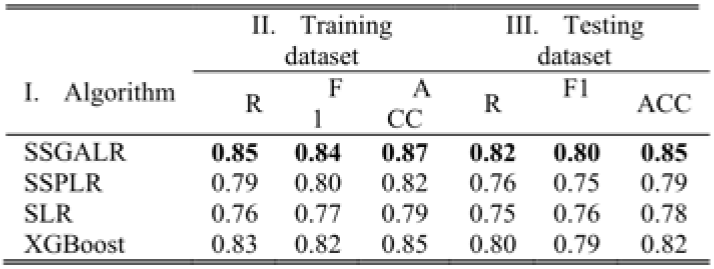
Overall, a layered approach using complementary techniques tailored to models, data and users is advised to realize the full potential of XAI. The spectrum ranges from inherently interpretable to flexible black box explanation methods.

Article: Semi-Supervised Learning Classification Based on Generalized Additive Logistic Regression for Corporate Credit Anomaly Detection

Published In: IEEEXplore

This article evaluates a semi-supervised learning classification model that uses generalized additive logistic regression (GALR) to detect corporate credit anomalies. Traditionally, companies have used supervised learning methods to evaluate corporate credit, but acquiring labeled sample data was often costly and time-consuming. The model evaluated in this article incorporates both labeled and unlabeled data, along with financial and non-financial variables. Among the four models tested—supervised semi-parametric logistic regression (SSPLR), supervised logistic regression (SLR), extreme gradient boosting (XGBoost), and GALR—GALR provided better variable selection and accuracy compared to the other models.

The GALR model was trained and tested on actual corporate data, including 3,584 companies, with 966 labeled samples and 2,618 unlabeled samples. Corporate default risk assessments most frequently use methods like neural networks, support vector machines, random forests, AdaBoost, and logistic regression models. Logistic regression models are the most widely used due to their explanatory power, prediction accuracy, and relatively simple calculations. However, they are parametric, including all available information, which can result in inaccurate risk predictors. This is why the authors proposed the generalized additive model (GAM) to achieve high risk prediction accuracy. GAMs simultaneously estimate model parameters and identify the selection of input variables, referred to as a semi-automatic additive model.



I. Algorithm	II. Training dataset			III. Testing dataset		
	R	F1	ACC	R	F1	ACC
SSGALR	**0.85**	**0.84**	**0.87**	**0.82**	**0.80**	**0.85**
SSPLR	0.79	0.80	0.82	0.76	0.75	0.79
SLR	0.76	0.77	0.79	0.75	0.76	0.78
XGBoost	0.83	0.82	0.85	0.80	0.79	0.82

GAMs encompass four general learning scenarios: clustering, classification, reduction, and regression. This method is widely used in applications such as facial recognition, image retrieval, and video segmentation.

To evaluate the model, financial variables such as solvency, development, and business classifications, cash flow, and profitability were used. Non-financial variables, such as customer satisfaction, national financial development, enterprise innovation classification, corporate governance, and corporate social responsibility, were also included. To preprocess the data, missing values were replaced with mean imputation, and outliers less than the 1% quantile and greater than the 99% quantile were replaced with the respective 1% quantile and 99% quartile. Using the bootstrap method, the data for the 3,584 companies was resampled 100 times to obtain the best recall value, F1 score, and accuracy. The analysis showed that variables such as the equity ratio, cash flow to debt ratio, equity multiplier, total assets turnover rate, net profit on total assets, net profit growth rate, number of penalties, credit of surrounding corporate entities, and risk transparency are highly indicative of credit anomalies. The findings underscore the value of non-financial data, especially for small and medium-sized businesses that may not have extensive financial data.

Article: General Pitfalls of Model-Agnostic Interpretation Methods for Machine Learning Models

Published In: xxAI - Beyond Explainable AI - International Workshop, Held in Conjunction with ICML 2020

Summary/Review:

This article focuses primarily on model-agnostic methods for IML (interpretable machine learning; it is worth noting that the authors use this term almost exclusively instead of xAI, though they seem interchangeable in this context). After providing a brief background on the need for and use cases of IML, the article provides an excellent summary of the major categories of IML methodologies. This is followed by an analysis of nine common pitfalls for practitioners to be aware of, as well as proposed solutions and a discussion of open issues. These categories are broken down into three categorizations based on source: unsuitable ML models, limitations of the IML method, and wrong application of the IML method.

The authors categorize IML models along two dimensions: local vs global interpretation methods, and feature effect vs feature importance. While local explanation methods focus on providing transparency around individual predictions or classifications, global models seek to "describe the expected behavior of the entire model with respect to the whole data distribution." While the authors state that their primary interest is in the pitfalls associated with global interpretation methods, they note that many of these pitfalls apply to local explanations as well (and in fact a considerable portion of the paper does still address these pitfalls with respect to local methods). Feature importance methods "quantify the contribution of a feature to the model performance… or to the variance of the prediction function," while feature effect indicates both "direction and magnitude of a change in predicted outcome due to changes in feature values."

To summarize, local methods describe the impact of a variable on a single prediction, while global methods describe the impact of a variable more generally across the entire model; feature importance describes the magnitude of a variable's influence on the model's prediction, while feature effects describe both the magnitude and direction of that influence. Methods of particular interest to our project include the local, feature effects methods of LIME (local interpretable model-agnostic explanations), counterfactual explanations, Shapley values, and SHAP (SHapley Additive explanation), as well as the global effects-based method of PDP (partial density plots) and global importance method of PFI (permutation feature importance).

This paper does an excellent job of providing examples and context for common mistakes practitioners can make using these various methods. While discussing the "One-Fits-All" pitfall, the authors explain that the manner in which each of these methods quantifies feature rank/importance is important to understanding which contexts they are suitable for. They point out as an example that feature importance measured by a loss-based method (such as PFI) is an excellent way to gain insight into the impact of a variable on how well the model generalizes on unseen data (the model's generalization error), but can be misleading if the goal is to identify which features a model relies upon to make its predictions. By contrast, a method such as SHAP importance better quantifies the relevance of a feature regarding a model's prediction. Another important concern raised in this context is that some methods can provide different explanations depending on the selected hyperparameters; the authors that that, "For counterfactuals, explanation goals are encoded in their optimization metrics such as sparsity and data faithfulness; The scope and meaning of LIME explanations depend on the kernel width and the notion of complexity."

Other discussed pitfalls that resonated with me include "Bad Model Generalizations" (essentially the idea that interpretability is only useful if your model describes the underlying data-generating process well), "Unnecessary Use of Complex Models" (we only truly need interpretability models when simpler, more transparent models to not meet our needs), and "Unjustified Causal Interpretation" (mistaking the importance of a feature to the model's decision-making with a causal relationship in the underlying data-generating process).

Overall, this was a lengthy and highly detailed paper that covered a lot of ground with regard to a wide variety of IML/xAI methodologies. It has provided some excellent insight that will help us to evaluate and select the techniques that will be most applicable and useful to our project. I also plan to save this article as a valuable reference for the future.

Article: Explainable Machine Learning for Fraud Detection

Published In: IEEE Computer Special Issue on Explainable AI and Machine Learning

This article evaluates two prominent explainable AI methods (Local Interpretable Model-Agnostic explanations, or LIME and SHapley Additive explanation, or SHAP) in the specific context of fraud detection. Both of these methods are considered attribution techniques, which explain a single response variable prediction by providing ranks for the predictor variables estimated to have been most important to that individual prediction.

By the authors' description, "LIME approximates the predictions of the underlying black box model by training local surrogate models to explain individual predictions. Essentially, LIME modifies a single data sample by tweaking the feature values in the simpler local model and observes the resulting impact on the output." By comparison, "The SHAP method explains the prediction of an instance by computing the contribution of each feature to the prediction using Shapley values based on coalition game theory. Intuitively, SHAP quantifies the importance of each feature by considering the effect each possible feature combination has on the output." I found these concise and approachable explanations very important, as being able to describe any technique employed simply and clearly is essential to the fundamental concept of explainable AI.

This study used the IEEE-CIS Fraud Detection dataset to compare 8 predictive models (naïve bayes, logistic regression, decision trees, gradient boosted trees, random forests, neural networks, autoencoders, and isolation forest). Both LIME and SHAP were then used to identify and rank the top 10 most important features of the same single instance prediction across all eight models. These explanations were compared to the global interpretation of the logistic regression model, which was treated as the ground truth due to "the transparency of the logistic regression model and its wide acceptance among regulatory bodies."

Overall, the authors found that LIME and SHAP tended to identify similar top features, though with different rankings among those variables. While the rankings of features varied across both evaluated methods and all eight models, the authors concluded that on average SHAP produces explanations closer to the established "ground truth" provided by the global explanations of the linear regression model.

Another important topic that the authors discuss is how to establish a background dataset for the SHAP method. While an all black image is considered a standard reference point in image classification, no such universal reference point exists for financial fraud data. The authors explore the use of two background datasets, one made up entirely of fraudulent transactions and the other entirely of normal transactions. They report that “models like Naive Bayes, Logistic Regression and Decision Trees give more consistent explanations regardless of the background dataset, while models like Random Forest, Gradient Boosting and Neural Networks are more sensitive to the reference point,” though they do note that all models exhibit some degree of sensitivity to the reference point. Additionally, for the real-time application of any of these models at scale, the potential size of the background data set can get incredibly large. Using the entire available background dataset may often be quite computationally expensive, while reducing the size has implications for the accuracy of the explanations. The authors state that selecting the appropriate background set “should be based on the goals of the explanation,” but do not elaborate any further.

Imbalanced data challenges and correction

Article: Performance evaluation of class balancing techniques for credit card fraud detection

Published In: 2017 IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI)

A primary concern across most fraud detection problems is data imbalance. Typically, instances of fraud tend to be relatively rare compared to genuine transactions, which creates a number of challenges for traditional machine learning and artificial intelligence models. This imbalance can be corrected in two primary ways: undersampling (essentially discarding a portion of majority class observations), or oversampling (generating realistic synthetic examples of the minority class).

This article conducted a thorough investigation of several methods for addressing class imbalance in credit card fraud data. The authors employed five methods of oversampling and four methods of undersampling, which were then evaluated across multiple classification models. Oversampling methods included four variations of synthetic minority over-sampling technique (SMOTE), as well as simple random oversampling with replacement. Undersampling methods consisted of three variations of condensed nearest neighbor (CNN), as well as simple, non-heuristic random under sampling. All seven methods were evaluated using a cost-sensitive decision tree model (C4.5), AdaBoost, and bagging. Additionally, a cost-sensitive support vector machine (csvm) was used to evaluate the four undersampling methods (though the authors do not mention why this was not also used on the oversampling methods).

Across the various oversampling methods, the Edited Nearest Neighbors variation of SMOTE performed the best across all three classification models (based on both AUC-ROC and G-Mean metrics), while SAFE SMOTE consistently performed the worst. I was also surprised to see that simple random over sampling performed reasonably well, particularly when combined with AdaBoost and bagging. AdaBoost and bagging appear to perform comparably well across the majority of balancing methods tested, while the cost-sensitive decision tree consistently performed worse than the other two. Undersampling in general appears to perform slightly worse compared to the better oversampling methods, though both the TL method and random undersampling only slightly underperformed compared to ENN SMOTE. The authors assert that the TL method produced the best results out of the undersampling methods, though upon inspection of the results I find it hard to argue that TL performed definitively better than random undersampling (especially after discarding the C4.5 classifier, which was the overall worst classification model). Given the incredible simplicity of random undersampling, I am surprised that this method was not given more attention for the results it achieved.

Overall this was a very thorough paper that covered a wide variety of methodologies relevant to our project. While it was useful to see a variety of class balancing techniques compared, I would have liked to have also seen the results from the original imbalanced dataset for comparison. I am also unsure how the authors came to the conclusion that the TL method of undersampling produced the best results, as simple random undersampling appears to have achieved very comparable results.

I find the results achieved with random undersampling particularly interesting, as this would allow for balancing the classes using only genuine, non-synthetic data. This could be a valuable tool for use with Explainable AI. The process is not only very simple to understand and explain, but reliance on 100% genuine data may also be perceived as more trustworthy and transparent in situations where concerns about bias may be present (as is the case with our project). It is not addressed in this paper, but it would be interesting to explore potential ways to conduct random undersampling that would still retain the majority of the information from the original dataset. In particular, I am intrigued by a possible approach to bagging where repeated random undersampling is conducted in an ensemble manner prior to the classification method being applied. This would increase the computational cost of the analysis, but potentially provide an increased level of trust and transparency.

References

1. A machine learning based credit card fraud detection using the GA algorithm for feature selection https://journalofbigdata.springeropen.com/articles/10.1186/s40537-022-00573-8
2. A supervised machine learning algorithm for detecting and predicting fraud in credit card transactions https://www.sciencedirect.com/science/article/pii/S2772662223000036#b29
3. Credit Card Fraud Detection using Artificial Neural Network and Back Propagation https://ieeexplore.ieee.org/abstract/document/9120957?casa_token=KPtXjWIV-H0AAAAA:EaF2-S4H5hho0EDCQYpgoWt68-SHWw_kpYLkfBPid2aVG1rexh4RLJUciVp42bVmnWLP3nbvDQ
4. Explainable AI: current status and future directions https://arxiv.org/pdf/2107.07045.pdf
5. Semi-Supervised Learning Classification Based on Generalized Additive Logistic Regression for Corporate Credit Anomaly Detection https://ieeexplore.ieee.org/document/9246557
6. General Pitfalls of Model-Agnostic Interpretation Methods for Machine Learning Models https://link.springer.com/chapter/10.1007/978-3-031-04083-2_4
7. Explainable Machine Learning for Fraud Detection https://ar5iv.labs.arxiv.org/html/2105.06314
8. Performance evaluation of class balancing techniques for credit card fraud detection https://ieeexplore.ieee.org/abstract/document/8392219?casa_token=N-GrdYOfXQkAAAAA:sMce_valhyxqxO2mhxgAbEsitrGYQHMd_kFTx8hLZOc97hvNLvdCYNPj1iPx4zDod7P_AoKSeg
9. Effective feature selection technique in an integrated environment using enhanced principal component analysis. https://doi.org/10.1007%2Fs12652-019-01647-x
10. Fraud detection in E-banking by using the hybrid feature selection and evolutionary algorithms. http://scholar.google.com/scholar_lookup?&title=Fraud%20detection%20in%20E-banking%20by%20using%20the%20hybrid%20feature%20selection%20and%20evolutionary%20algorithms&journal=Int%20J%20Comput%20Sci%20Netw%20Secur&volume=17&issue=8&pages=271-279&publication_year=2017&author=Pouramirarsalani%2CA&author=Khalilian%2CM&author=Nikravanshalmani%2CA
11. Explainable AI - Understanding and Trusting Machine Learning Models https://www.datacamp.com/tutorial/explainable-ai-understanding-and-trusting-machine-learning-models