Evolution of Credit Scoring Models and Explainable AI Techniques in Loan Defaulter Prediction

Samar Patil
samar.patil5@mail.dcu.ie
School of Computing,
Dublin City University
23260190

Shree Sudame
shree.sudame2@mail.dcu.ie
School of Computing,
Dublin City University
23263322

I. Introduction

Loans are essential for borrowers, covering needs like mortgages, auto financing, personal expenses, education, and small business. For lenders, loans generate revenue and enhance consumer relationships, enabling the offering of diverse monetary services. To minimize loan losses, financial institutions assess repayment ability using disbursal conduct and credit scores. Approval or rejection is based on financial and credit history data, with credit scores representing repayment ability. Lenders categorize clients into high-risk and low-risk, favouring low-risk borrowers with positive credit scores [1]. Manual processing of numerous daily loan applications is time-consuming and resource-intensive, prone to inaccuracies [2]. To address these issues, banks are increasingly adopting machine learning algorithms for efficient application processing, improved customer satisfaction, and accurate assessment of customer creditworthiness. Credit scoring models are widely used in banking for loan assessments, online retail for credit risk evaluation, and insurance, especially in auto insurance, for predicting claims and assessing premium reliability.[4]

In the realm of data-driven decision-making in financial institutions, our analysis will centre on key components of the credit scoring model like examining different datasets, optimizing feature engineering, addressing class imbalances, evaluating metrics, and refining modelling strategies. Furthermore, we will explore Explainable AI (XAI) algorithms to address transparency and interpretability issues within the credit scoring framework.

1) Explainable Artificial Intelligence: Financial institutions grapple with the opacity of machine learning algorithms in credit scoring models [10]. Explainable Artificial Intelligence(XAI) algorithms address the issues for the black box nature of many AI models that can help them to plan arrangements and adaptations of policies accordingly thus preventing the potential risk. This concept was introduced by DARPA in 2016 to ensure accountability and trustworthiness in AI decision making process. The applications of XAI are immense ranging from Healthcare, Finance and Law. XAI can transform finance by enhancing understanding of AI decision-making and provide crucial insights to the customers as well as the institution to understand the factors behind why there loans were not sanctioned [16]. Financial institutions in credit scoring must ensure transparency to avoid biases [17]. Explanation for false positives is crucial as misclassifying bad credit as good credit poses significant risks for lenders [7].

II. Literature review

The evolution of credit scoring models reflects a continuous effort to enhance predictive performance and address the challenges associated with identifying high-risk customers.

A. Evolution of Credit Scoring Models

Spoorthi et al., 2021[2] conducted a comparative analysis of algorithms like Support Vector Machine (Linear, Gaussian, and Polynomial)

and Naive Bayes (Gaussian, Multinomial, and Bernoulli), as well as an ensemble classifier (Random Forest) to compare the prediction performance of the base and ensemble classifiers in the context of credit scoring. The Researchers' aim was to identify and understand various general parameters utilized by the banks while evaluating customer's profile during the loan application process. The researchers performed the study on the loan prediction dataset sourced from Kaggle. The results evaluated based on performance measure like Precision, Recall, F1, and Accuracy indicated that L-SVM outperformed other classifiers with an accuracy of 83.24%, alongside the Random Forest Classifier with 82.7%. Notably, the ensemble strategy effectively addresses imbalanced datasets, however, it overlooks class imbalance mitigation in other models as well as hyperparameter tuning, which could enhance model performance. Similar research was conducted by **Safiya, Parvin and Saleena., 2020[3]**, on the Australian Credit dataset. The researchers performed a comparative analysis among a range of base and ensemble classifiers, including Logistic Regression, Decision Tree, Support Vector Machine, K-Nearest Neighbour, Multi-Layer Perceptron, Nave Bayes, Random Forest, Ada Boosting, Gradient Boosting, Extra Tree Classifier, alongside a Voting ensemble of (Logistic Regression +Decision Tree +Support Vector Machine), aiming to identify the optimal model for credit scoring that accurately classifies high-risk applicants. The researchers performed dimensionality reduction using Principal Component Analysis (PCA) on the train and test sets for converting the 14-dimensional feature space to a 2-dimensional feature space which was later fed to all the classifiers. Using similar evaluation settings as [2] their study revealed that among base classifiers, Support Vector Machine outperformed with an accuracy of 87.68%, while among ensemble classifiers, Random Forest Classifier, Extra Tree Classifier, and Bagged Decision Tree outperformed with accuracies of 88.41%, 86.38%, and 88.41%, respectively. A positive aspect of the research was the inclusion of a broader range heterogeneous base and ensemble classifiers. However, there were some drawbacks. Additionally, standardization was not performed, potentially leading to poorer generalization. The study also lacked hyperparameter tuning, and the issue of class imbalance was not addressed.

Chen et al., 2022[4] proposed an advanced layered weighted and majority voting framework to develop a single generalized model which would outperform other baseline approaches implemented in the context of credit scoring, the researchers ranked 7 classifiers (Naive Bayes, Support Vector Machine, Extra Tree, Logistic Regression, K-Nearest Neighbor, Random Forest, and Decision Tree) on the German Loan Approval and Australian Credit Scoring datasets from the UCI Machine Learning Repository. The top 5 classifiers, selected based on f1-score, underwent further enhancement with AdaBoost (ADB) to improve overall predictive performance. ADB trained the ranked classifiers, acting as weak base learners on the same training set, to create an ultimate classifier. The model comprised two layers: the first layer integrated ADB_5, ADB_4, ADB_3 with majority voting, and the second layer integrated ADB_2, ADB_1 using majority voting. Feature selection methods such as STEP, CORR, t-test, NRS were employed for optimal results, and class imbalance was addressed using **SMOTE** oversampling. Z-score Normalization (Standardization) was applied to bring the data to a unit scale. The results displayed fine optimization under AdaBoost+LMV for the German dataset with an accuracy of **74.33%**, while STEP+AdaBoost+LMV outperformed others with an accuracy of **88.94%** in the Australian dataset. A positive aspect was the evaluation of model performance using metrics like Sensitivity, Specificity, G-measure, and ROC_AUC scores. However, the study did not address the generalization issue, as in layered weighted and majority voting frameworks, individual models are trained independently without continuous adaptation to data. Despite the identified drawbacks, the proposed model can demonstrate efficient performance in workplace settings similar to that in the research. In contrast to the previously suggested Layered Weighted Ensemble approach, **Chornous and Nikolskyi., 2018[5]** proposed a Layered Stacking Ensemble. This ensemble incorporates K-Nearest Neighbor, Decision Tree, and Support Vector Machine as input classifiers. These

classifiers are trained to generate outputs, which are then used as inputs for the Gradient Boosting (GB) classifier. Gradient Boosting serves as the primary (mainstream) classifier, focusing on binary classification to distinguish between defaulters and non-defaulters. The research aimed to provide a cost-effective solution suitable for small-scale businesses and lending institutions with limited operational budgets. The researchers concentrated on hybrid feature selection, utilizing a combination of techniques such as mean decrease Gini, information gain, and Chi-squared coefficient methods in a Majority Voting setting. This approach was employed to derive combined features for the hybrid model. The Hybrid feature selection strategy resulted in an **8%** improvement in model accuracy. The final accuracy of the Stacking Ensemble model, with the combined features, reached **84.1%**. A notable aspect of the paper is its intention to benefit small financial institutions with constrained computational resources. However, it's important to note that the model's limitation lies in its use of common and easily understandable algorithms. This decision may have been influenced by the researchers' emphasis on transparency, interpretability, or ease of implementation.

In a study by **Jabeen, Singh, and Vats., 2023[14]**, they addressed the class imbalance issue in credit scoring. The aim was to prevent misclassification of bad credit applicants and minimise financial risks for Credit Card Fraud Detection (CCFD) using SMOTE, which generated synthetic data points for the minority class (fraudulent transactions). Machine Learning techniques, specifically Decision Trees, Logistic Regression, and Random Forest were utilized. The study employed the European Cardholders dataset, which was highly imbalanced, with only 0.172% of transactions labelled as fraudulent. The study concluded that SMOTE significantly improved the performance of machine learning models in detecting fraudulent transactions. Random Forest emerged as the most effective algorithm across all performance metrics when used with SMOTE. The decision tree's accuracy, precision, recall, and F1-score improved with SMOTE compared to without SMOTE, demonstrating its effectiveness. In a similar study by **Parekh et al., 2021[13]**, obstacles such as class imbalance were effectively addressed by implementing Resampling Techniques including SMOTE and Random Under-Sampling. The investigation revealed that Random Forest with SMOTE yielded the most favourable outcomes, achieving a notable ROC-AUC score of 0.97. **Zhong and Wang., 2023[15]** focused on deep learning credit scoring models in Internet finance, highlighting the scarcity of research considering class imbalance issues. The proposed model combined Deep Forest (DF) with five resampling methods: random over-sampling (**ROS**), random under-sampling (**RUS**), synthetic minority oversampling technique (**SMOTE**), tomek links, and SMOTE+Tomek. Through experiments on four Internet financial credit scoring datasets, the RUS-DF model demonstrated its effectiveness among others.

Wei et al., 2021[6] aimed to address the limitations of cold start models in credit scoring as a common problem associated with the context is the lack of sufficient data samples making it difficult for the models to generalize. Thus the researchers proposed a transfer learning approach by implementing two models: the SPY-transfer model and the SPY-TrAdaBoost model. The SPY algorithm was transferred to the Positive Unlabeled (PU) field to select valuable samples from the source data for integration into the target data. The implementation involved a migration strategy that included relabeling and mixing the source field (negative samples) and target field (positive samples) into train and test sets. The study was performed on the credit scoring dataset published by the Qianhai Credit Information "Good Letter Cup" algorithm competition. The trained classifier then predicted the probability of each sample in the test set belonging to positive samples. A threshold was set, and samples with probabilities greater than the threshold were considered positive, expanding the target field dataset. Results demonstrated effective optimization under the SPY-TrAdaBoost + Cold Start (XGB) combination, showcasing a **2%** increase in the F1 score compared to the cold start model. This approach outperformed the Cold start model, achieving an impressive accuracy of **92.1%**. A notable positive aspect was the model's efficiency, especially in scenarios with a

limited number of experimental samples. However, a potential drawback emerged. There is a risk that the transfer-based approach might overfit the source field dataset, leading to generalization issues. Additionally, if biases present in the source field data are not representative of the target field data, the model may produce inaccurate or unfair predictions for specific groups of people.

B. Explainable AI

Explainable AI (XAI) as discussed in a study by **G. P. Reddy and Y. V. P. Kumar., 2023[16]** addresses concerns about the black-box nature of AI models by making them more interpretable and transparent. This concept was introduced by DARPA in 2016 to ensure accountability and trustworthiness in AI decision-making processes. Unlike conventional models, XAI methods help inferencing results of learned models. Methods used in XAI are feature visualization, saliency mapping and model interpretation. Model interpretation is a model-agnostic technique as it can be implemented regardless of the specific algorithm or model used and analyses the relationship between input features and model output. LIME (Local Interpretable Model-Agnostic Explanations) and SHAP (SHapley Additive exPlanations) are popular model-agnostic techniques that provide explanations for individual predictions or analyze global feature importance.

In their study, **Kotrachai et al., 2023[12]** implemented the SHapley Additive exPlanations (SHAP) method to enhance the model interpretability of credit card fraud detection systems utilizing machine learning models like K-Nearest Neighbors (KNN), Random Forests, Extreme Gradient Boosting (XGBoost), and Logistic Regression. The researchers generated SHAP summary plots for each of the machine learning models employed in the study. Then, a comparative evaluation before and after feature selection was conducted based on their SHAP values, ensuring that only the most informative features were retained for further analysis. Feature V15 was identified as the most influential, alongside other significant features. For instance, a trend was observed in the KNN model, where Feature 15 again showed the highest impact (+0.15), followed by Feature 13 (+0.08), and Feature 5 (+0.04), thus concluding the significant role of Feature 15 in predicting credit card fraud. Despite some metrics decreasing post-feature selection, models maintained high precision scores, indicating robustness. However, while the paper provided a commendable comparative evaluation of different models in credit card fraud detection, it could have further enhanced its contribution by expanding the comparison to include a wider array of state-of-the-art techniques and algorithms. Additionally, a more comprehensive assessment of performance metrics beyond the standard ones, such as recall and ROC-AUC would have provided a more holistic view of model effectiveness.

A similar study by **Mahajan and Shukla., 2023[11]** utilized the SHAP framework to analyze false positive cases, providing an in-depth analysis of how specific features contributed and how interactions among these features influenced bankruptcy prediction. The study addressed imbalance using Class weighing techniques on Taiwans bankruptcy dataset and the Polish company dataset, generating both local and global explanations. The initial experiment revealed that borrowing dependency was the most crucial feature for false positive cases, particularly for companies classified as bankrupt. Meanwhile, Net Income to Total Assets emerged as the most significant feature in determining the overall model feature importance. In a second experiment, researchers observed varying orders of feature importance. Given the differences in feature importance for false positive cases, SHAP waterfall plots were used to analyze individual company explanations further. Features with large positive or negative values strongly influenced predictions, as demonstrated by SHAP interaction values. Techniques used like class weights and HalvingGridSearchCV might not fully address the issue of imbalance and flexibility to define the search space and objective functions respectively.

XAI has vast applications in areas like Healthcare, Law, and Finance, especially in improving credit scoring for precise lending decisions and reducing risks. However, there's a challenge in balancing between explainability and performance. Furthermore, the lack of a common standard makes it challenging to develop, stressing the importance of essential evaluation metrics on what constitutes

a good explanation [16].

III. CONCLUSIONS

As the number of loan applicants is expected to increase in the coming years, both customers and financial institutions will demand accurate loan assessment models for precise risk management. This review focuses on predicting high-risk loan applicants, addressing class imbalance with techniques like SMOTE and Random Under-Sampling [13][14][15] to enhance model performance. Extensive comparative analysis, conducted in [3][4], emphasizes that models like Logistic Regression and Support Vector Machine (SVM) outperform other base approaches, while ensemble models like Random Forest and AdaBoost (ADB) outperform among ensemble approaches. The review specifically highlights Explainable AI (XAI), focusing on techniques such as LIME and SHAP [10][11], to improve the interpretability and explainability of credit scoring models. These techniques provide extensive feedback for defaulted applicants and personalized loan recommendation suggestions. Our practicum aims to bridge research gaps in the interpretability and transparency of black-box credit scoring models. This involves implementing XAI techniques, enhancing model performance through a more robust optimization framework, and reducing risks through extensive classifier comparisons.

REFERENCES

[1] Y. Zhao, "Credit Card Approval Predictions Using Logistic Regression, Linear SVM and Nave Bayes Classifier," 2022 International Conference on Machine Learning and Knowledge Engineering (MLKE), Guilin, China, 2022, pp. 207-211, doi: 10.1109/MLKE55170.2022.00047. keywords: Analytical models;Machine learning algorithms;Statistical analysis;Computational modeling;Static VAr compensators;Support vector machine classification;Predictive models;machine learning;credit scoring;classification;Logistic Regression;Linear SVM,

[2] Spoorthi, B. *et al.* (2021) Comparative Analysis of Bank Loan Defaulter Prediction Using Machine Learning Techniques, in *2021 IEEE International Conference on Distributed Computing, VLSI, Electrical Circuits and Robotics (DISCOVER)*. *2021 IEEE International Conference on Distributed Computing, VLSI, Electrical Circuits and Robotics (DISCOVER)*, Nitte, India: IEEE, pp. 2429. Available at: https://doi.org/10.1109/DISCOVER52564.2021.9663662.

[3] Safiya Parvin, A. and Saleena, B. (2020) An Ensemble Classifier Model to Predict Credit Scoring - Comparative Analysis, in *2020 IEEE International Symposium on Smart Electronic Systems (iSES) (Formerly iNiS)*. *2020 IEEE International Symposium on Smart Electronic Systems (iSES) (Formerly iNiS)*, Chennai, India: IEEE, pp. 2730. Available at: https://doi.org/10.1109/iSES50453.2020.00017.

[4] Chen, R., Ju, C.H. and Shen.Tu., F. (2022) A Credit Scoring Ensemble Framework using Adaboost and Multi-layer Ensemble Classification, in *2022 4th International Conference on Pattern Recognition and Intelligent Systems*. *PRIS 2022: 2022 4th International Conference on Pattern Recognition and Intelligent Systems*, Wuhan China: ACM, pp. 7279. Available at: https://doi.org/10.1145/3549179.3549199.

[5] Chornous, G. and Nikolskyi, I. (2018) Business-Oriented Feature Selection for Hybrid Classification Model of Credit Scoring, in *2018 IEEE Second International Conference on Data Stream Mining Processing (DSMP)*. *2018 IEEE Second International Conference on Data Stream Mining Processing (DSMP)*, Lviv: IEEE, pp. 397401. Available at: https://doi.org/10.1109/DSMP.2018.8478534.

[6] Wei, Q., Liu, Y. and Wu, K. (2021) Transfer Learning Based Credit Scoring, in *2021 IEEE 24th International Conference on Computer Supported Cooperative Work in Design (CSCWD)*. *2021 IEEE 24th International Conference on Computer Supported Cooperative Work in Design (CSCWD)*, Dalian, China: IEEE, pp. 12511255. Available at: https://doi.org/10.1109/CSCWD49262.2021.9437749.

[7] Barua, S. *et al.* (2021) Swindle: Predicting the Probability of Loan Defaults using CatBoost Algorithm, in *2021 5th International Conference on Computing Methodologies and Communication (ICCMC)*. *2021 5th International Conference on Computing Methodologies and Communication (ICCMC)*, Erode, India: IEEE, pp. 17101715. Available at: https://doi.org/10.1109/ICCMC51019.2021.9418277.

[8] Jafar, A. and Lee, M. (2023) HypGB: High Accuracy GB Classifier for Predicting Heart Disease With HyperOpt HPO Framework and LASSO FS Method, *IEEE Access*, 11, pp. 138201138214. Available at: https://doi.org/10.1109/ACCESS.2023.3339225.

[9] Moe, S.T. and Nwe, T.T. (2023) A Hybrid Approach of Logistic Regression with Grid Search Optimization in Credit Scoring Modeling for Financial Institutions, in *2023 IEEE Conference on Computer Applications (ICCA)*. *2023 IEEE Conference on Computer Applications (ICCA)*, Yangon, Myanmar: IEEE, pp. 6266. Available at: https://doi.org/10.1109/ICCA51723.2023.10181624.

[10] Sharma, V. and Midhunchakkaravarthy, D. (2023) XGBoost Classification of XAI based LIME and SHAP for Detecting Dementia in Young Adults, in *2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT)*. *2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, Delhi, India: IEEE, pp. 16. Available at: https://doi.org/10.1109/ICCCNT56998.2023.10307791.

[11] Mahajan, A. and Shukla, K.K. (2023) Analyzing False Positives in Bankruptcy Prediction with Explainable AI, in *2023 International Conference on Artificial Intelligence and Applications (ICAIA) Alliance Technology Conference (ATCON-1)*. *2023 International Conference on Artificial Intelligence and Applications (ICAIA) Alliance Technology Conference (ATCON-1)*, Bangalore, India: IEEE, pp. 15. Available at: https://doi.org/10.1109/ICAIA57370.2023.10169390.

[12] Kotrachai, C. *et al.* (2023) Explainable AI supported

Evaluation and Comparison on Credit Card Fraud Detection Models, in *2023 7th International Conference on Information Technology (InCIT)*. *2023 7th International Conference on Information Technology (InCIT)*, Chiang Rai, Thailand: IEEE, pp. 8691. Available at: https://doi.org/10.1109/InCIT60207.2023.10413100.

[13] Parekh, P. *et al.* (2021) Credit Card Fraud Detection with Resampling Techniques, in *2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT)*. *2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, Kharagpur, India: IEEE, pp. 17. Available at: https://doi.org/10.1109/ICCCNT51525.2021.9579915.

[14] Jabeen, U., Singh, K. and Vats, S. (2023) Credit Card Fraud Detection Scheme Using Machine Learning and Synthetic Minority Oversampling Technique (SMOTE), in *2023 5th International Conference on Inventive Research in Computing Applications (ICIRCA)*. *2023 5th International Conference on Inventive Research in Computing Applications (ICIRCA)*, Coimbatore, India: IEEE, pp. 122127. Available at: https://doi.org/10.1109/ICIRCA57980.2023.10220646.

[15] Zhong, Y. and Wang, H. (2023) Internet Financial Credit Scoring Models Based on Deep Forest and Resampling Methods, *IEEE Access*, 11, pp. 86898700. Available at: https://doi.org/10.1109/ACCESS.2023.3239889.

[16] G. P. Reddy and Y. V. P. Kumar, "Explainable AI (XAI): Explained," 2023 IEEE Open Conference of Electrical, Electronic and Information Sciences (eStream), Vilnius, Lithuania, 2023, pp. 1-6, doi: 10.1109/eStream59056.2023.10134984.

[17] X. Dastile, T. Celik and H. Vandierendonck, "Model-Agnostic Counterfactual Explanations in Credit Scoring," in IEEE Access, vol. 10, pp. 69543-69554, 2022, doi: 10.1109/ACCESS.2022.3177783.