Explainable and Fair Credit Default Prediction Across Multiple Financial Datasets Using Machine Learning

Divyanshi Singh
Department of Computer Science and Engineering
Manipal University Jaipur
Jaipur, India
divyanshi.sg8@gmail.com

***Abstract*—Accurate prediction of consumer credit defaults is crucial for reducing financial risk and promoting responsible lending practices. This study introduces a machine learning framework that prioritizes explainability and fairness for predicting credit defaults, assessed through three publicly available datasets that reflect various financial contexts. The models, including Logistic Regression, Random Forest, and XGBoost, are trained following suitable preprocessing, feature engineering, and, when needed, strategies for handling class imbalance. The effectiveness of the models is assessed using ROC-AUC, cross-validation, and tuning thresholds focused on recall to improve the identification of high-risk customers. To enhance explainability, SHAP is applied to pinpoint significant risk factors influencing model outcomes, and risk-based segmentation is conducted based on predicted probabilities. Experimental results show that XGBoost achieves the best trade-off between discrimination and minority-class recall across datasets, with ROC-AUC values of approximately 0.77 on the Taiwan credit card dataset, 0.80 on the German credit dataset, and 0.86 on the Give Me Some Credit dataset. Optimizing thresholds on a large-scale dataset greatly enhances the recall for defaulters and uncovers the swap between precision and recall. The research underscores variations in feature significance across different datasets and stresses the need for transparency and careful tuning related to risk in practical credit scoring systems.**

***Index Terms*—Credit risk, default prediction, explainable AI, fairness, SHAP, XGBoost, class imbalance, threshold tuning**

I. Introduction

Credit risk assessment constitutes a fundamental element of contemporary financial decision-making systems. Financial institutions increasingly utilize predictive models to estimate the probability of borrower default, facilitating reductions in financial losses, more accurately calibrated lending policies, and enhanced systemic stability. Traditional credit scoring methods have predominantly employed statistical techniques such as logistic regression due to their simplicity and regulatory approval; however, these models frequently fail to capture complex nonlinear relationships inherent in large, heterogeneous financial datasets.

As large-scale credit and behavioral data have grown, machine learning models like Random Forest and Gradient Boosting have shown better predictive performance than traditional methods. Still, many complex models are hard to understand, which is a problem in finance where transparency, fairness, and accountability matter. Regulators and ethical standards now expect models to explain individual decisions and prevent unintended bias against demographic groups.

Furthermore, a significant portion of the current literature assesses models based on a singular dataset and prioritizes accuracy at the expense of interpretability and fairness. In contrast, this study examines model performance and behavior in three different credit environments, integrating explainable AI approaches with demographic fairness assessment and cost-based evaluation. Additionally, threshold tuning is utilized to explore how model behavior can be adjusted to emphasize the identification of high-risk customers in settings sensitive to risk.

This study advances traditional research on explainable credit scoring by incorporating fairness mitigation, threshold tuning, and cross-dataset validation into a comprehensive predictive framework. In contrast to previous studies that primarily emphasize model performance or post-hoc explanation, this research integrates explainability, fairness intervention, and business-oriented risk simulation to develop responsible and practical credit decision systems. The integration of demographic bias mitigation, local explanation case studies, and cost-based evaluation further enhances transparency and operational relevance.

The main contributions of this paper are:

- An analysis comparing various machine learning models using three publicly accessible credit risk datasets.
- Incorporating SHAP-based explanation techniques to pinpoint key risk factors on both global and local scales.
- A threshold-based sensitivity analysis to improve recall for defaulters under different operating points.
- An analysis of fairness across different demographic age groups, examining the effects of fairness mitigation on both bias metrics and the estimation of financial risk.

II. Problem Statement and Research Questions

While machine learning models can deliver impressive results in predicting credit defaults, there are still various obstacles to address. Initially, models developed on a specific dataset may not perform effectively in different credit situations. Additionally, the opaque nature of black-box predictions diminishes transparency and hampers confidence in automated decision-making systems. Finally, demographic variables may affect model results, leading to potential fairness issues. This study addresses the following research questions:

- How well do machine learning models generalize across different credit datasets?
- Which financial features consistently influence default prediction?
- Can threshold tuning improve detection of high-risk customers?
- Does model prediction vary significantly across age groups?

III. Related Work

Traditional credit scoring models have predominantly employed Logistic Regression due to its interpretability and regulatory acceptance. Recent research, however, indicates that ensemble models, such as Random Forest and Gradient Boosting, surpass classical statistical methods in terms of predictive performance. Explainable AI techniques, particularly SHAP, have been increasingly utilized to interpret black-box models within financial applications. Furthermore, fairness-aware machine learning has emerged as a significant research domain, concentrating on the identification and mitigation of demographic biases in automated decision-making systems.

Nonetheless, the majority of existing studies assess models using a single dataset and primarily emphasize performance metrics. Few investigations compare model behavior across multiple datasets while incorporating explainability and fairness analysis. This study endeavors to address this gap.

IV. Datasets

A. Taiwan Credit Card Dataset

The initial dataset is the widely recognized Taiwan credit card default dataset, comprising 30,000 customer entries and 25 original variables, which include an identifier, demographic information, credit limits, billing amounts, and repayment status over a period of six months. After eliminating the identifier column, 24 features are left. The binary target variable *default payment next month* denotes whether the customer failed to make their payment in the subsequent month. The distribution of classes is imbalanced, featuring 23,364 clients who did not default and 6,636 who did.

B. German Credit Dataset

The second dataset comprises the numeric representation of the German Credit data, which includes 1,000 instances and 24 feature columns along with a target label. The features reflect various socio-economic and credit-related attributes such as account status, credit amount, loan duration, employment, and personal background. The original target values are transformed into a binary label, wherein *1* indicates good credit and *2* corresponds to the default (bad credit) category.

C. Give Me Some Credit Dataset

The third dataset utilized is the Give Me Some Credit dataset, which includes nearly 150,000 records of individual borrowers, with 30,000 samples set aside for testing purposes in this analysis. The target variable *SeriousDlqin2yrs* indicates whether an individual faced significant financial hardship (for instance, being 90 days overdue) within a two-year period. The input features encompass revolving utilization of unsecured credit lines, age, counts of overdue events across various time frames, debt ratio, monthly income, the total number of open credit lines and loans, the total number of real estate loans or lines, and the number of dependents. Missing values in *MonthlyIncome* and *NumberOfDependents* are replaced with the median prior to the training of the model.

V. Methodology

A. Preprocessing and Feature Engineering

For the Taiwan dataset, the identifier column is removed and several domain-inspired features are engineered to capture aggregate behaviour:

- **UTIL_RATIO**: ratio of the most recent bill amount to the credit limit.
- **TOTAL_BILL**: sum of bill amounts over the past six months.
- **TOTAL_PAY**: sum of payment amounts over the past six months.
- **PAY_SCORE**: mean of the six payment status variables (PAY_0–PAY_6), representing overall delinquency severity.
- **BILL_TO_LIMIT**: ratio of total bill across six months to the credit limit.

These engineered features are appended to the original variables. For all datasets, the feature matrix X is constructed by dropping the target column, and the target vector y is defined as the corresponding default indicator.

For the Give Me Some Credit dataset, missing values in *MonthlyIncome* and *NumberOfDependents* are filled with the respective median values. For the German Credit data, numeric columns are used directly without additional feature engineering.

For models sensitive to scale (Logistic Regression), features are standardized using *StandardScaler* fitted on the training set and applied to the test set.

B. Train–Test Split and Class Imbalance

Each dataset is divided into training and test sets through stratified sampling to maintain the class distribution. An 80% training and 20% testing split is typically employed.

In the case of the Taiwan dataset, the Synthetic Minority Oversampling Technique (SMOTE) is utilized to rebalance the classes during training. For the Give Me Some Credit dataset, the significant imbalance in *SeriousDlqin2yrs* is addressed by employing the *scale_pos_weight* parameter in XGBoost to increase the weight of the minority class.

C. Models

Three supervised learning models are evaluated on each dataset:

- **Logistic Regression**: a linear baseline classifier trained with a maximum of 1000 iterations on standardized features.
- **Random Forest**: An ensemble comprising 200 decision trees, characterized by default depth and inherent randomness, was trained on the original feature space.
- **XGBoost**: gradient boosted trees with tuned hyperparameters. For the Taiwan dataset, the tuned model uses 800 estimators, maximum depth of 8, learning rate of 0.02, subsample and column subsample of 0.9, minimum child weight of 4, gamma of 0.1, and *scale_pos_weight* of 3.5. For the German and Give Me Some Credit datasets, more compact configurations with 300 estimators and depth 6 are used, and class weights are adjusted where appropriate.

D. Evaluation Metrics and Cross-Validation

The performance of the model is evaluated on the test sets through accuracy, precision, recall, F1-score, and ROC-AUC metrics. Given that default prediction is characterized by imbalanced classification, special focus is placed on the recall and F1-score for the default class, along with ROC-AUC as a measure independent of thresholds.

Furthermore, for the Taiwan dataset, stratified 5-fold cross-validation is utilized with ROC-AUC as the evaluation metric to determine model stability. The optimized XGBoost model attains cross-validation ROC-AUC scores in the range of 0.77 to 0.78, with a mean close to 0.774, which reflects a reliable discriminative performance.

E. Explainability with SHAP

For the top-performing tree-based models on each dataset, SHapley Additive exPlanations (SHAP) are utilized to assess the contributions of features. Tree-based explainers are used to calculate SHAP values for the corresponding test sets, and summary plots are created to illustrate global feature significance and the nature of their influence.

F. Risk Segmentation and Threshold Tuning

On the Taiwan dataset, predicted default probabilities from XGBoost are used to define three risk levels:

- Low risk: predicted probability < 0.30
- Medium risk: $0.30 \leq p < 0.60$
- High risk: $p \geq 0.60$

Risk levels are designated to test samples, and their distribution is depicted to show how customers are categorized according to model results.

For the Give Me Some Credit dataset, threshold tuning is applied to the XGBoost probabilities to study the trade-off between recall and precision for defaulters. In addition to the default threshold of 0.5, alternative thresholds of 0.30 and 0.20 are evaluated.

VI. Results and Performance Evaluation

Across various datasets, XGBoost consistently demonstrated superior performance. The ROC-AUC values were approximately 0.77 for the Taiwan dataset, 0.80 for the German dataset, and 0.86 for the large-scale credit distress dataset. Cross-validation confirmed the model's stability and generalizability.

Threshold tuning revealed a significant impact on recall. At the default threshold, recall for defaulters was moderate. Lowering the threshold substantially improved recall, facilitating better identification of high-risk individuals, albeit at the expense of reduced precision and accuracy.

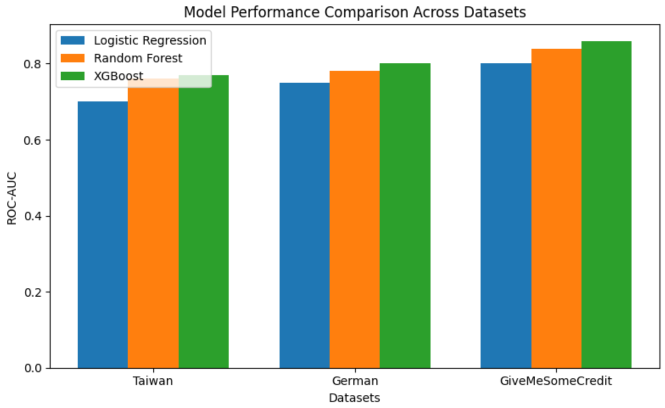

Fig. 1. ROC-AUC comparison of Logistic Regression, Random Forest, and XGBoost across the Taiwan, German Credit, and Give Me Some Credit datasets.

VII. Explainability Analysis

For each dataset, SHAP values are calculated utilizing tree-based explainers on the respective XGBoost models to offer both global and local interpretability.

In the case of the Taiwan dataset, SHAP summary plots reveal that recent payment status variables (such as PAY_0 and PAY_2), the engineered PAY_SCORE feature, and aggregated billing and payment quantities (TOTAL_BILL, TOTAL_PAY, BILL_TO_LIMIT, and UTIL_RATIO) are among the most significant predictors of default. Elevated values of delinquency scores and higher utilization or bill-to-limit ratios are associated with an increased predicted risk of default, whereas higher total payments generally correlate with a decreased risk.

In the German Credit dataset, the SHAP analysis identifies features related to credit amount, loan duration, and certain account-status indicators as significant factors influencing default prediction. Increased credit amounts and extended loan durations correlate with a heightened risk, while favorable account status codes and shorter loan durations are associated with a reduced likelihood of default.

In the Give Me Some Credit dataset, SHAP values indicate that the utilization of revolving unsecured credit lines, debt-to-income ratio, and the frequency of past-due events (30–59 days, 60–89 days, and 90+ days) are crucial risk factors.

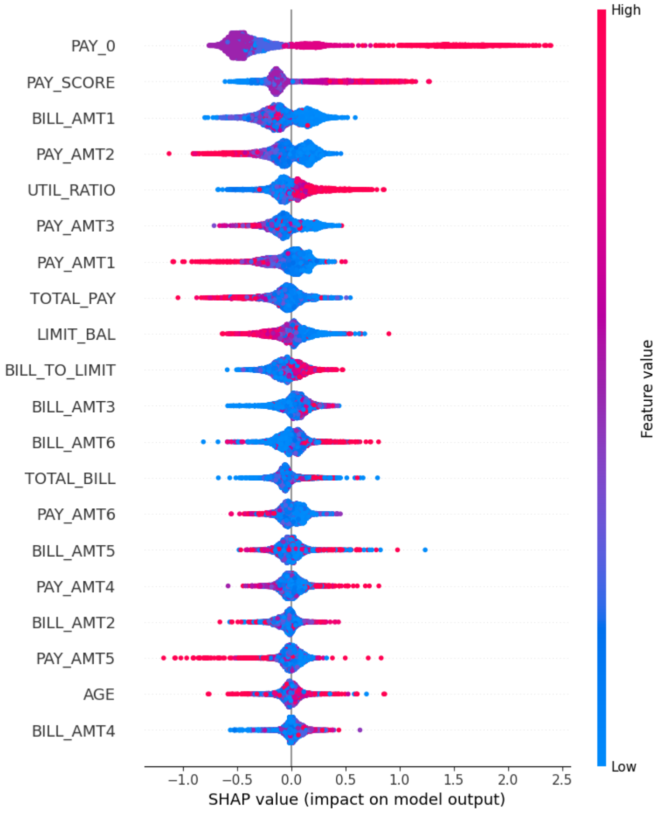

Fig. 2. SHAP summary plot for the Taiwan credit card dataset using the tuned XGBoost model. Recent payment status and utilization-related features are the most influential predictors of default.

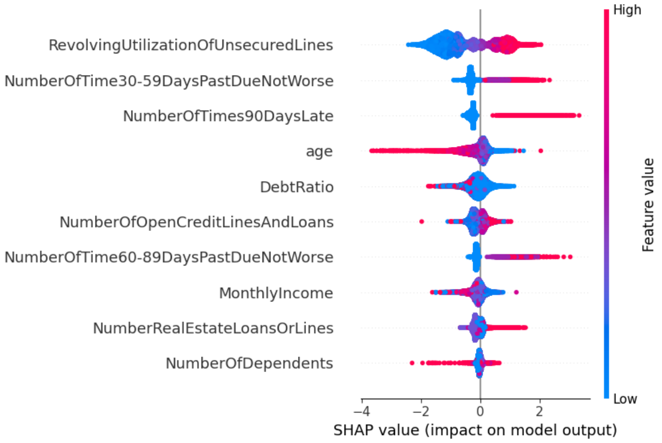

Fig. 3. SHAP summary plot for the Give Me Some Credit dataset. Revolving utilization, debt ratio and past-due event counts are the most important drivers of financial distress.

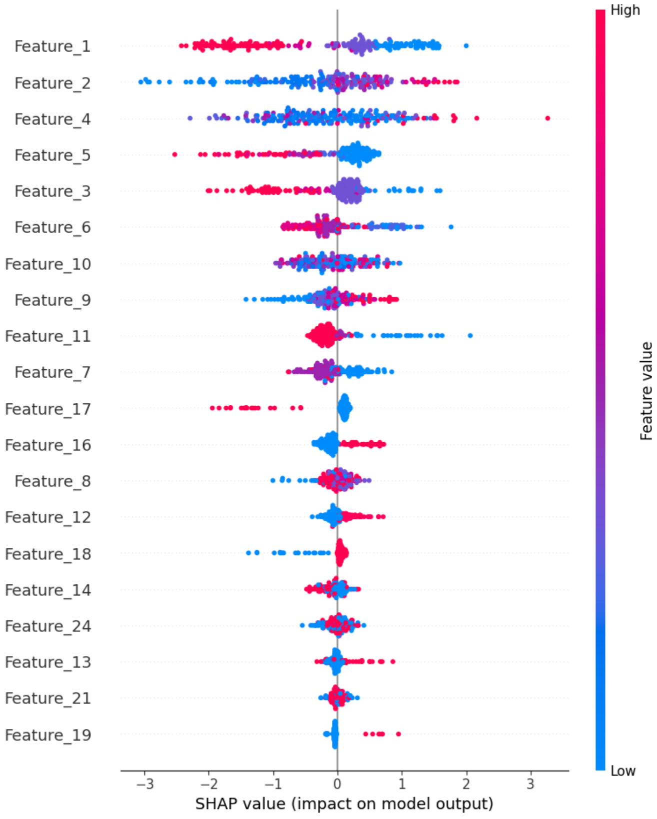

Fig. 4. SHAP summary plot for the German Credit dataset. Credit amount, loan duration and account status features dominate the model's default predictions.

Customers exhibiting very high utilization rates and numerous past-due occurrences receive considerably higher SHAP contributions toward the risk of default. Monthly income and the quantity of open credit lines also contribute as secondary influences, affecting risk when considered alongside past due history.

The findings indicate that the XGBoost models are capturing significant financial trends that resonate with expert understanding in every dataset, and that SHAP offers a clear breakdown of individual predictions by illustrating the contributions of each feature.

A. *Local SHAP Case Study*

To further elucidate individual-level interpretability, a local SHAP explanation was generated for a borrower predicted to have a very high default risk. The SHAP force plot indicates that delayed payment statuses (PAY 0, PAY 2, PAY 5) and a high aggregated delinquency score (PAY SCORE) are the most significant factors contributing to the prediction of default. Furthermore, the absence of recent payments in specific months (PAY AMT3 and PAY AMT6) exacerbates the predicted risk. These findings are consistent with domain intuition, as persistent repayment delays and low payment activity are recognized indicators of financial distress. This case study illustrates the capability of SHAP to provide actionable, human-readable explanations for individual credit decisions and supports the practical application of the model in risk review workflows.

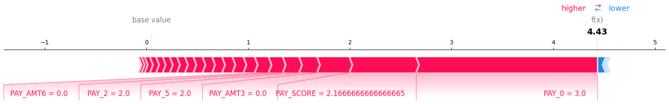

Fig. 5. SHAP force plot explaining an individual high-risk borrower prediction.

VIII. Fairness Analysis

A demographic examination across various age categories revealed that the anticipated risk of default diminishes as age increases. The younger age groups exhibited higher predicted risk levels, consistent with the patterns of default observed, but also suggested that the model might enhance demographic inequalities. This highlights the necessity of reviewing machine learning models for equity in financial contexts.

A. *Fairness Mitigation Results*

A fairness mitigation strategy utilizing reweighted training was implemented to address demographic bias across age groups. Prior to mitigation, the demographic parity difference was approximately 0.015, while the equal opportunity difference was around 0.029. Following the application of fairness-aware reweighting, these values were reduced to approximately 0.010 and 0.022, respectively, indicating a significant reduction in prediction disparity while preserving overall model performance. This outcome suggests that fairness interventions can enhance equity in risk estimates without significantly compromising predictive accuracy, and it demonstrates the feasibility of fairness-aware modeling in practical credit scoring contexts.

TABLE I
Fairness and Risk Metrics Before and After Mitigation

Metric	Before	After
Demographic parity difference	0.015	0.010
Equal opportunity difference	0.029	0.022
Estimated financial risk	3728	3678

IX. Discussion

The findings indicate that ensemble models deliver robust predictive capabilities in various credit scenarios. An analysis of explainability offers important perspectives on risk factors, while adjusting thresholds enables the model to be tailored for different risk management approaches. Nevertheless, the research underscores the compromises between accuracy and recall, as well as the necessity to ensure demographic fairness is observed.

A. *Cost-Based Business Impact Analysis*

A cost-based assessment was performed to evaluate the operational effects of fairness-aware modeling within a scenario of asymmetric loss where the penalties for missed defaulters are more significant than for false positives. Utilizing this straightforward risk model, the fairness-mitigated classifier lowered the projected financial risk from 3728 to 3678 units. This decrease suggests that the integration of fairness-aware learning can enhance both equity and risk management outcomes. In other words, the mitigation process not only addresses demographic bias but could also result in more financially sound lending decisions in practical applications.

X. Limitations and Future Work

This study uses publicly available datasets and does not include real-time financial data. Future work could explore cost-sensitive learning, fairness-aware optimization, and deeper feature engineering to improve performance and reduce bias.

XI. Conclusion

This document introduces a framework for predicting credit defaults that emphasizes both explainability and fairness across various datasets. The findings indicate that machine learning models, especially XGBoost, perform well while also offering interpretable insights via SHAP. Adjusting thresholds enhances the identification of high-risk individuals, and the analysis of fairness underscores the necessity of responsible AI in financial decisions. The suggested method shows excellent generalization and practical significance for systems assessing credit risk.

References

[1] I.-C. Yeh and C.-H. Lien, “The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients,” *Expert Systems with Applications*, vol. 36, no. 2, pp. 2473–2480, 2009.

[2] S. Lessmann, B. Baesens, H.-V. Seow, and L. C. Thomas, “Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research,” *European Journal of Operational Research*, vol. 247, no. 1, pp. 124–136, 2015.

[3] S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2017, pp. 4765–4774.

[4] M. T. Ribeiro, S. Singh, and C. Guestrin, “Why should I trust you?: Explaining the predictions of any classifier,” in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining (KDD)*, 2016, pp. 1135–1144.

[5] T. Chen and C. Guestrin, “XGBoost: A scalable tree boosting system,” in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining (KDD)*, 2016, pp. 785–794.

[6] G. Lemaître, F. Nogueira, and C. K. Aridas, “Imbalanced-learn: A Python toolbox to tackle the curse of imbalanced datasets in machine learning,” *Journal of Machine Learning Research*, vol. 18, no. 17, pp. 1–5, 2017.

[7] S. A. Aarfi, M. Khan, and S. M. Danish, “Predicting credit card default using machine learning,” *American Journal of Information Science and Technology*, vol. 13, no. 1, pp. 7–17, 2024.

[8] F. Wahab, S. Khan, and M. A. Raza, “Credit card default prediction using machine learning and deep learning techniques,” *Machine Learning with Applications*, vol. 16, 100521, 2024.

[9] N. Kaur and S. Kumar, “Prediction of credit card defaults through data analysis and machine learning techniques,” *Materials Today: Proceedings*, vol. 51, pp. 2250–2256, 2022.

[10] T. K. Reddy and S. Gupta, "Predicting credit card defaults with machine learning," *International Journal for Research in Applied Science and Engineering Technology*, vol. 11, no. 9, pp. 1475–1482, 2023.

[11] K. Mangla, A. N. Srivastava, and M. R. Gupta, "SHAP and LIME: An evaluation of the discriminative power in credit risk," *Frontiers in Artificial Intelligence*, vol. 4, 2021, Art. no. 752558.

[12] L. Lin and Y. Wang, "SHAP stability in credit risk management: A case study in credit card default model," *arXiv preprint* arXiv:2508.01851, 2025.

[13] S. Barocas, M. Hardt, and A. Narayanan, *Fairness and Machine Learning*. Cambridge, MA, USA: fairmlbook.org, 2019.

[14] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2016, pp. 3315–3323.

[15] UCI Machine Learning Repository, "Default of Credit Card Clients Data Set," 2016. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients

[16] UCI Machine Learning Repository, "Statlog (German Credit Data)," 2023. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data)

[17] Kaggle, "Give Me Some Credit," 2011. [Online]. Available: https://www.kaggle.com/competitions/GiveMeSomeCredit