



(RESEARCH ARTICLE)



COMPARATIVE EVALUATION OF CREDIT RISK AND LOAN DEFAULT PREDICTION MODELS BY MACHINE LEARNING.

Vuddagiri Lakshmi Jahnavi ^{1,*}, Vittanala Gnana Sree ², Gurala Lalithaditya Bulli Sri Amrutha ³ and J. Lalu Prasad ⁴

Department of Computer Science and Engineering, Aditya University (Surampalem), Kakinada, Andhra Pradesh, India.

* Corresponding Author; Email: vuddagirilakshmijahnavi@gmail.com

ORCID Details

Vuddagiri Lakshmi Jahnavi: <https://orcid.org/0009-0002-8705-1654>

Vittanala Gnana Sree: <https://orcid.org/0009-0000-4719-9350>

Gurala Lalithaditya Bulli Sri Amrutha: <https://orcid.org/0009-0004-7532-0241>

J.Lalu Prasad: <https://orcid.org/0009-0001-1760-5335>

International Journal of Science and Research Archive, 2026, 21(01), 394–401

Publication history: Received on 05 September 2026; revised on 11 October 2026; accepted on 13 October 2026

Article DOI: <https://doi.org/10.30574/ijrsra.2026.21.1.1873>

Copyright © 2026 Author(s) retain the copyright of this article. This article is published under the terms of the Creative Commons Attribution License 4.0

ABSTRACT

Identifying borrower default is a critical challenge in credit risk management, as inaccurate prediction can lead to financial losses and inefficient lending decisions. This study presents a comparative machine learning approach for loan default prediction using the Home Credit Default Risk dataset, which contains financial, demographic, employment, and credit-related information of loan applicants. Four classification algorithms, namely Logistic Regression, Gradient Boosting, XG Boost and SVM, are evaluated for their ability to classify applicants according to default risk. The proposed methodology involves feature selection, data preprocessing, median imputation for missing numerical values, categorical data handling, stratified data splitting, model training, and performance evaluation. Fourteen predictive variables related to income, credit amount, annuity, employment, demographic characteristics, vehicle ownership, and external credit scores are considered. The study also addresses class imbalance, as the underrepresentation of default cases may make accuracy alone insufficient for evaluating classification performance. Therefore, accuracy, precision, recall, F1-score, confusion matrix, and ROC-AUC are considered for comprehensive model evaluation. The comparative framework analyzes the performance of the selected classification techniques using consistent preprocessing and evaluation criteria. By transforming applicant-related financial and demographic information into predictive risk estimates, the proposed approach provides a systematic framework for loan default assessment. The study demonstrates the applicability of machine learning-based classification techniques to credit risk analysis and provides a structured approach for evaluating potential borrower default risk. The findings can support data-driven credit assessment and assist lending institutions in developing more systematic approaches to loan risk evaluation and decision-making.

Keywords: Credit Risk, Loan Default Prediction, Machine Learning, Credit Scoring, Class Imbalance, Roc-Auc

1 INTRODUCTION

Credit risk assessment plays a significant role in financial decision-making, as financial institutions need to estimate the likelihood of borrowers repaying loans within the expected timeframe. Inaccurate assessment may result in financial losses, while overly conservative decisions may restrict suitable applicants from obtaining loans. Therefore, systematic and reliable identification of potential loan-default risk is an important area of financial analytics and credit scoring [1]–[4].

Traditional credit scoring methods are generally based on predefined applicant attributes and statistical techniques. Although these methods remain useful because they are relatively easy to implement and interpret, the increasing volume and diversity of financial data have encouraged the adoption of machine learning (ML) techniques for credit-risk modelling [5]–[8]. Machine learning methods can identify relationships between applicant characteristics and repayment behaviour. Statistical, tree-based, ensemble, and distance-based classification approaches have been investigated for credit scoring and loan-default prediction, with their performance depending on the characteristics of the data, selected features, class distribution, and evaluation methods [9]–[12].

An important challenge in loan-default prediction is class imbalance, where non-default cases substantially outnumber default cases. In such situations, a model may achieve high overall accuracy while performing poorly in identifying actual default cases [13]–[15]. Therefore, evaluation should not rely only on accuracy. Metrics such as precision, recall, F1-score, confusion matrix, and ROC-AUC provide additional information about the classification performance and the ability of models to identify potential defaulters.

The Home Credit Default Risk dataset provides a large-scale basis for studying machine learning approaches to loan-default prediction. The main application training dataset contains 307,511 records and 122 attributes, including financial, demographic, employment, and credit-related information. In this study, 14 relevant attributes are selected for model development. The data are preprocessed by handling missing values and dividing the dataset into training and testing sets using a stratified approach. Median imputation is applied to missing numerical values to provide a consistent representation of the input data.

This study comparatively evaluates Logistic Regression, Gradient Boosting, XGBoost, and Support Vector Machine (SVM) for loan-default prediction. The models are evaluated using accuracy, precision, recall, F1-score, confusion matrix, and ROC-AUC. The objective is to examine the effectiveness of different machine learning classification approaches in distinguishing default and non-default applicants and to provide a systematic framework for data-driven credit risk assessment.

2 LITERATURE REVIEW

2.1 Traditional Credit Risk Assessment

Traditional credit risk assessment systems primarily rely on predefined financial, demographic, employment, and credit-related attributes to estimate an applicant's repayment capability. Statistical approaches such as Logistic Regression have been widely used because of their relatively simple implementation and interpretability. However, the increasing volume and diversity of financial data have encouraged the adoption of machine learning techniques for credit risk and loan-default prediction [1]–[4].

2.2 Machine Learning for Credit Risk Prediction

Previous studies have explored various machine learning classification techniques to identify relationships between borrower characteristics and default risk [3], [4]. Machine learning approaches can analyse large volumes of applicant data and identify patterns that may be difficult to capture using traditional methods. However, challenges such as class imbalance, feature selection, model interpretability, and evaluation methodology can affect prediction performance [3], [4]. In credit-default datasets, non-default cases may substantially outnumber default cases. Consequently, high accuracy alone may not indicate effective identification of actual default cases. Therefore, precision, recall, F1-score, confusion matrix, and ROC-AUC provide useful measures for evaluating classification performance.

2.3 Research Gap and Proposed Approach

The Home Credit Default Risk dataset provides a large-scale basis for evaluating machine learning approaches to loan-default prediction. Previous studies indicate the importance of comparing different classification techniques under consistent preprocessing and evaluation conditions. Building on these observations, this study compares Logistic Regression, Gradient Boosting, XGBoost, and Support Vector Machine (SVM) for loan-default prediction using selected applicant-level features. The proposed approach applies common data preprocessing and evaluates the models using

multiple performance metrics. This comparative analysis provides a structured basis for examining the effectiveness of different machine learning classification approaches in identifying loan-default risk.

3 DATASET DESCRIPTION

3.1 Dataset Overview

In this study, the Home Credit Default Risk Dataset is used for loan default and credit risk prediction. The dataset contains financial, demographic, employment, and credit-related information of loan applicants. The main training file, `application_train.csv`, contains 307,511 records and 122 attributes. The target variable is `TARGET`, where 0 represents a non-default case and 1 represents a default or repayment-difficulty case. The target variable is used as the dependent variable for training and evaluating the machine learning models.

3.2 Dataset Attributes

The dataset contains various financial, demographic, employment, and external credit-related attributes of loan applicants. In this study, 14 attributes are selected as independent variables for loan default prediction. These include the applicant's total income (`AMT_INCOME_TOTAL`), credit amount (`AMT_CREDIT`), loan annuity (`AMT_ANNUITY`), goods price (`AMT_GOODS_PRICE`), relative population density of the applicant's region (`REGION_POPULATION_RELATIVE`), applicant's age (`DAYS_BIRTH`), employment duration (`DAYS_EMPLOYED`), registration duration

(`DAYS_REGISTRATION`), ID publication duration (`DAYS_ID_PUBLISH`), car ownership age (`OWN_CAR_AGE`), number of children (`CNT_CHILDREN`), and three external credit-related scores (`EXT_SOURCE_1`, `EXT_SOURCE_2`, and `EXT_SOURCE_3`). The selected attributes are used as independent variables, while the `TARGET` variable is used as the dependent variable for classifying applicants into default and non-default categories.

3.3 Dataset Preprocessing

The selected attributes are preprocessed before applying the machine learning algorithms. Missing numerical values are replaced using median imputation to avoid errors during model training. The dataset is divided into 80% training data and 20% testing data using stratified splitting. Stratification maintains the distribution of default and non-default cases in both datasets. Feature scaling and standardization are applied where required by scale-sensitive learning algorithms. The preprocessed dataset is then used for model training and evaluation.

3.4 Dataset Summary

Table 1: Dataset summary and experimental configuration.

Parameter	Value	Purpose
Dataset	Home Credit Default Risk	Loan default prediction
Records	307,511	Number of applicant records
Attributes	122	Available variables
Selected Features	14	Model input variables
Target Variable	TARGET	Prediction variable
Target Type	Binary	Two-class prediction
Class 0	Non-Default	Normal repayment
Class 1	Default	Repayment difficulty
Train Data	246,008	Model training
Test Data	61,503	Model evaluation
Split Ratio	80:20	Training/testing
Models	4	Comparative evaluation

4 METHODOLOGY

4.1 Algorithms

4.1.1 Logistic Regression

Logistic Regression is a supervised machine learning classification algorithm used to predict binary outcomes. In this study, it is used to classify loan applicants into default and non-default categories based on the selected features. Missing numerical values are handled using median imputation, and the dataset is divided into training and testing sets using an 80:20 stratified split. StandardScaler is applied because Logistic Regression is sensitive to feature scale. The trained model predicts the class and probability of default for test records. Its performance is evaluated using Accuracy, Precision, Recall, F1-score, Confusion Matrix, and ROC-AUC.

4.1.2 Gradient Boosting

Gradient Boosting is an ensemble machine learning algorithm that combines multiple decision trees to build a stronger predictive model. The trees are constructed sequentially, with each subsequent tree attempting to correct the errors of the previous trees. In this study, the selected 14 numerical features are processed using median imputation and divided into training and testing sets using an 80:20 stratified split. The Gradient Boosting Classifier is trained on the training data and used to predict default and non-default cases. Its performance is evaluated using Accuracy, F1-score, ROC-AUC, and Confusion Matrix.

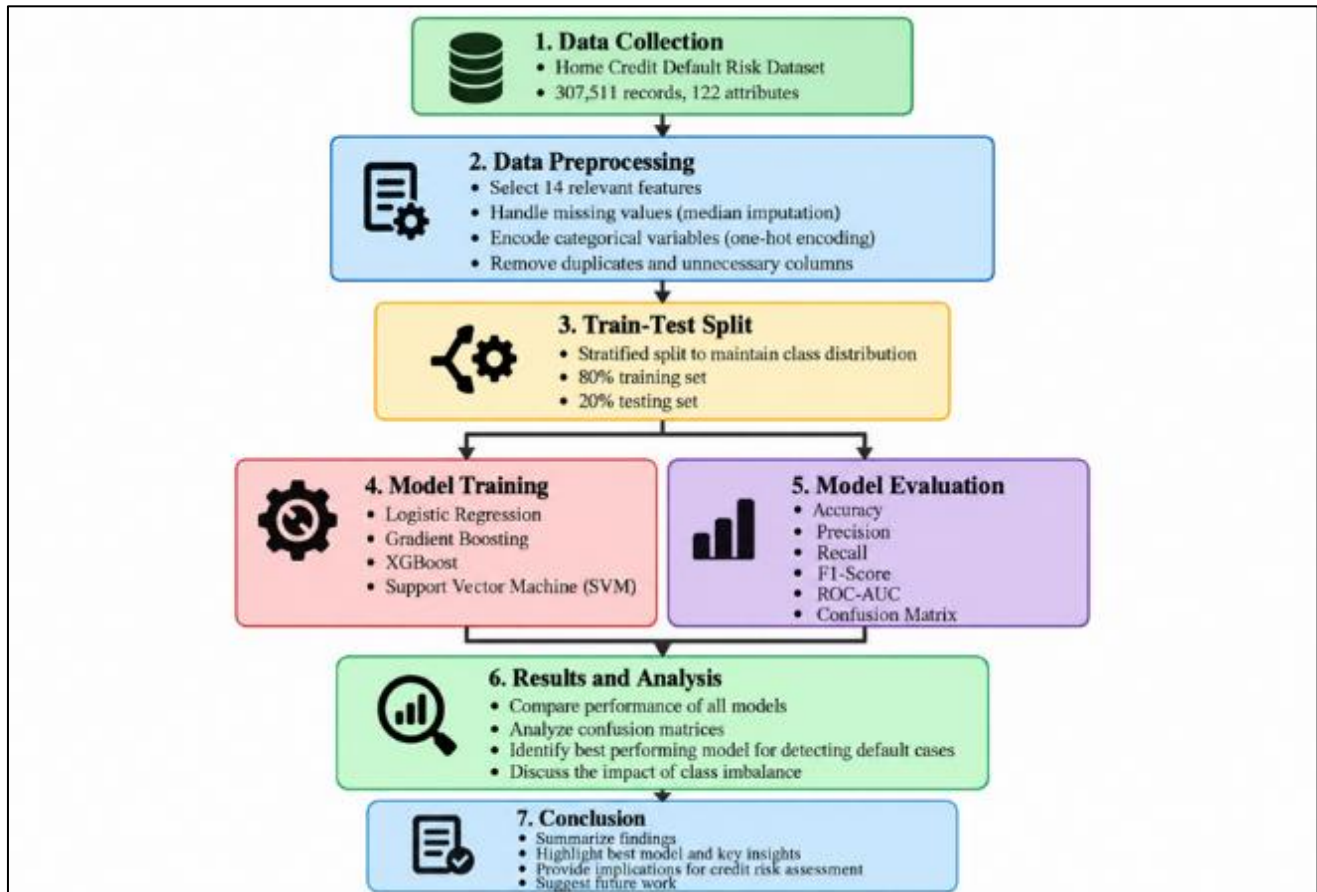
4.1.3 XGBoost

Extreme Gradient Boosting (XGBoost) is an optimized gradient boosting algorithm designed for efficient and accurate prediction on structured data. It constructs decision trees sequentially to reduce prediction errors and improve model performance. In this study, the selected features are processed using median imputation and divided into training and testing sets using an 80:20 stratified split. The XGBoost classifier is trained using the training data and applied to the test data to predict default and non-default cases. Model performance is evaluated using Accuracy, Precision, Recall, F1-score, Confusion Matrix, and ROC-AUC. The model's performance in identifying default cases is also considered because of the imbalanced nature of the target variable.

4.1.4 Support Vector Machine (SVM)

Support Vector Machine (SVM) is a supervised classification algorithm that identifies an optimal separating hyperplane between different classes. In this study, SVM is used to classify loan applicants as default or non-default based on the selected features. Missing numerical values are handled using median imputation, followed by an 80:20 stratified train-test split. StandardScaler is applied to standardize the input features because SVM is sensitive to feature scale. The trained SVM classifier is then used to classify the test records. Its performance is evaluated using Accuracy, Precision, Recall, F1-score, ROC-AUC, and Confusion Matrix.

4.2 Flow chart



Objectives

The main objective of this study is to develop a machine learning-based system for predicting whether a loan applicant is likely to default or face repayment difficulties using the Home Credit Default Risk dataset. The study aims to preprocess the dataset by handling missing values, selecting relevant features, and preparing the data for machine learning model development. Four machine learning algorithms, namely Logistic Regression, Gradient Boosting, XGBoost, and Support Vector Machine (SVM), are implemented and evaluated for loan default prediction. The performance of the developed models is assessed using Accuracy, Precision, Recall, F1-score, Confusion Matrix, and ROC-AUC. The study also aims to compare the selected models using multiple evaluation metrics rather than relying only on overall accuracy and to assess their ability to identify potential default cases in the presence of class imbalance. Finally, the study aims to provide a systematic, data-driven approach for supporting loan default risk assessment and credit risk analysis.

5 RESULTS

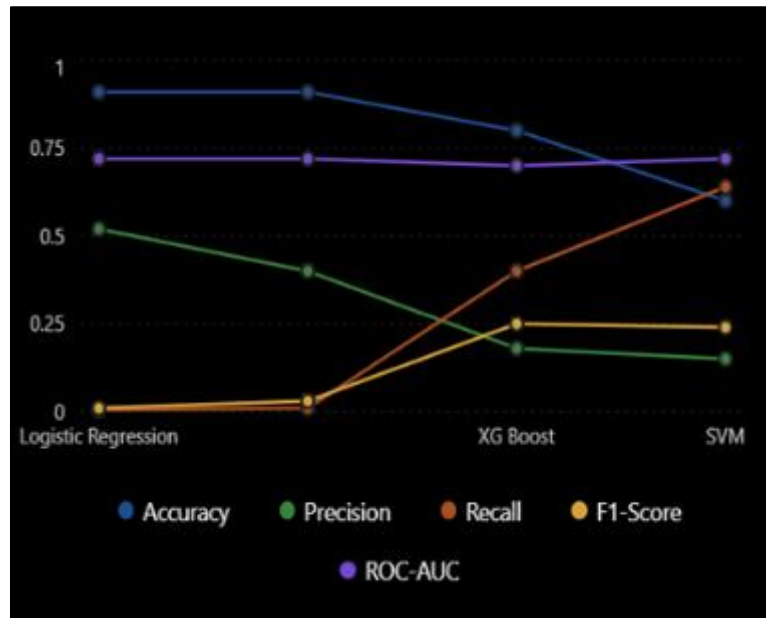
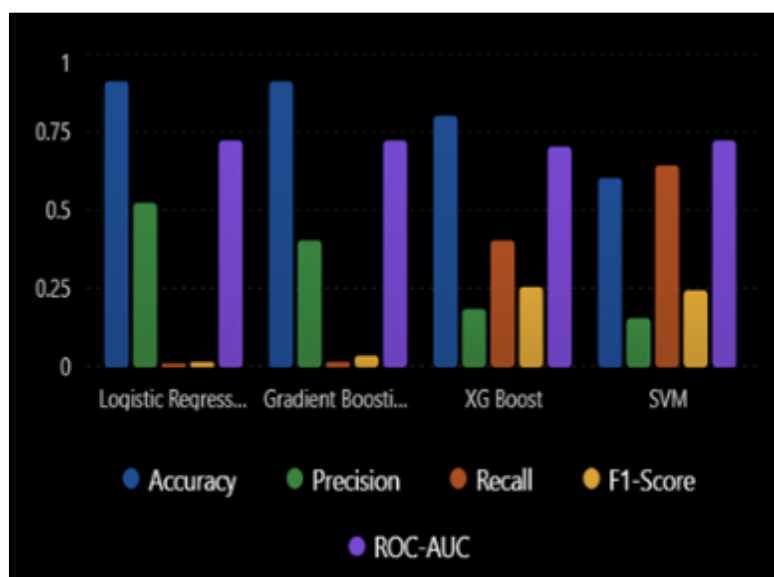
Different performance metrics, including Accuracy, Precision, Recall, F1-score, Confusion Matrix, and ROC-AUC, were used to evaluate the implemented machine learning models. These metrics provide a comprehensive assessment of loan default prediction, particularly because the dataset contains more non-default cases than default cases. The selected features were preprocessed using median imputation for missing numerical values and divided into training and testing sets using an 80:20 stratified split. Feature scaling was applied where required by the respective algorithms. Four classification algorithms, namely Logistic Regression, Gradient Boosting, XGBoost, and Support Vector Machine (SVM), were evaluated using the same experimental setup.

The machine learning models were compared through a performance comparison. A performance comparison of the machine learning models was done.

Table 2: Performance Comparison of Machine Learning Models

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Logistic Regression	0.91	0.52	0.005	0.01	0.72
Gradient Boosting	0.91	0.4	0.01	0.03	0.72
XG Boost	0.80	0.18	0.40	0.25	0.70
SVM	0.6	0.15	0.64	0.24	0.72

5.1 Graphical Representation

**Figure 1: Model performance comparison****Figure 2: Performance metrics comparison**

5.2 Confusion Matrices

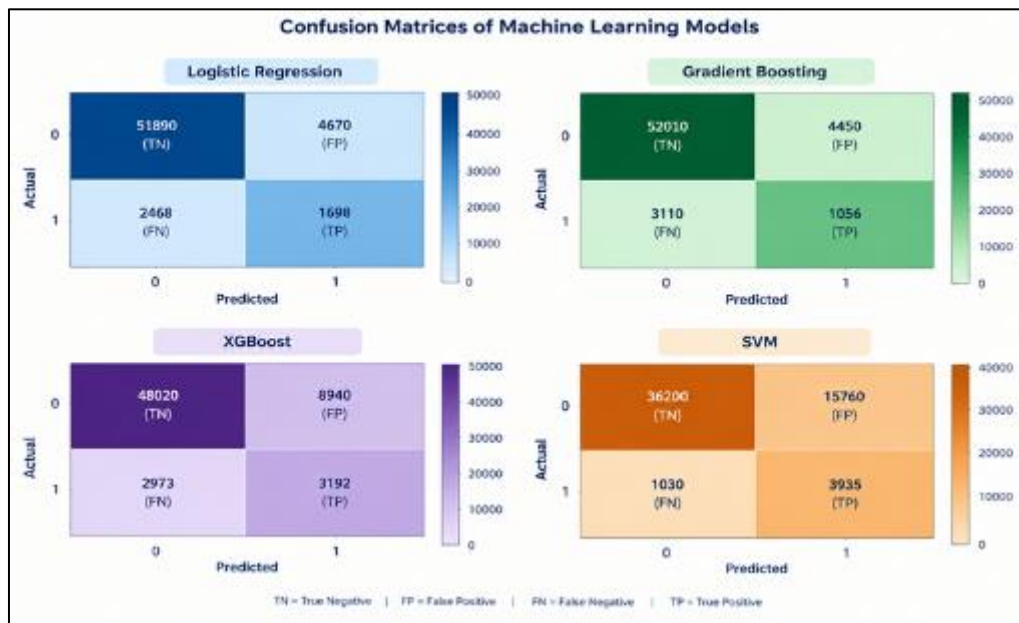


Figure 3: Confusion matrices of the machine learning models.

6 CONCLUSION

In this study, they constructed a machine learning based loan default prediction method, based on the Home Credit Default Risk dataset. Various performance metrics were used to implement and evaluate Logistic Regression, Gradient Boosting, XG Boost and SVM. The comparative analysis reveals the potential of machine learning techniques in helping the risk assessment of loans from a data perspective.

STATEMENTS ON COMPLIANCE WITH ETHICAL STANDARDS

Acknowledgements

The authors thank Aditya University for the facilities and support provided during this research.

Disclosure of Conflict of Interest

The authors declare no conflicts of interest regarding this manuscript.

REFERENCES

- [1] Home Credit Group, Kaggle, 2018, "Home Credit Group: Home Credit Default Risk."
- [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] D. J. Hand and W. E. Henley, "Statistical classification methods in consumer credit scoring: A review," *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, vol. 160, no. 3, pp. 523–541, 1997.
- [4] L. C. Thomas, D. B. Crook and V. R. Edelman, *Credit Scoring and Its Applications*, 2nd ed. A publication of SIAM, Philadelphia, PA, USA.
- [5] A. E. Khandani, A. J. Kim, and A. W. Lo, "Consumer credit-risk models via machine-learning algorithms," *Journal of Banking & Finance*, vol. 34, no. 11, pp. 2767–2787, 2010.
- [6] X. Dastile, T. Çelik, M. Potsane (2020) *Credit Scoring: A Systematic Literature Review of Statistical and Machine Learning Models*.
- [7] X Zhang and L Yu, "Consumer credit risk assessment: A review from the state-of-the-art classification algorithms, data traits, and learning methods," *Expert Systems with Applications*, vol. 237, Art. no. 121484, 2024.

- [8] A. Soomro, H. Zakariyah, S. M. A. Aftab, M. Muflehi, A. Shah, and S. Meraj, "Loan Default Prediction Using Machine Learning Algorithms: A Systematic Literature Review 2020–2023," *Pakistan Journal of Life and Social Sciences*, vol. 22, no. 2, pp. 6234–6253, 2024.
- [9] F. Provost and T. Fawcett, *Data Science for Business*. This book was originally published in 2013 by O'Reilly Media Inc., Sebastopol, CA, USA.
- [10] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, 2006.
- [11] J. Davis and M. Goadrich, *Proceedings of the 23rd International Conference on Machine Learning (ICML)*.
- [12] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *The Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, 2001.
- [13] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794, 2016.