International Journal on Science and Technology (IJSAT)
E-ISSN: 2229-7677 ● Website: www.ijsat.org ● Email: editor@ijsat.org

Improving Credit Score Classification Using Predictive Analysis and Machine Learning Techniques

Rahul M. Koshti [1], Prof. Shailesh J Molia [2], Prof. (Dr) Dhaval J Varia [3]

[1] Master's in Information Technology, Vishwakarma Government Engineering College, Gujarat, India
[2, 3]Asso. Prof. in Information Technology, Vishwakarma Government Engineering College, Gujarat, India

Abstract

Credit score classification is crucial to financial decision-making, allowing institutions to evaluate creditworthiness and manage risk accurately. Statistical models, on which conventional credit scoring techniques are often based, have difficulty capturing subtle patterns in credit information. This research examines the performance of various ML models such as Decision Trees, Logistic Regression, SVM, Random Forest, LightGBM, Gradient Boosting (GB), AdaBoost, XGBoost and Naïve Bayes (NB) on three benchmark datasets: Australian Credit dataset, German Credit dataset, and Kaggle Credit Score Classification dataset. All the models are run through feature selection, hyperparameter tuning to achieve improved prediction accuracy. Experimental findings reveal that Random Forest performs significantly better than all other models with the highest accuracy of 90.58% on the Australian Credit dataset, and XGBoost performed next best but at a slightly lower level than RF model. The above findings are indicative of the power of ensemble learning methods in credit risk assessment and offer a solid model for data-driven financial decision-making and mitigating lending risks for banks.

Keywords: Credit Score Classification, Machine Learning, Credit Risk, Predictive Analysis, Feature Selection, Model Optimization.

1. Introduction

In today's fast-paced and globalized economy, credit serves as one of the pillars to enable financial operations and support economic growth. Financial institutions and lenders apply credit scoring models to evaluate the credit worthiness of businesses, individuals, and entities applying for financial services [1]. A credit score, in the form of a number, facilitates the determination of the risk involved in extending credit, allowing lenders to make effective decisions. Consequently, credit scoring has been a critical application in the banking industry, with implications for the availability of credit and the financing prospects of borrowers. The advancement of credit scoring models over the years has been driven by progress in data analysis, statistical methodologies, and access to financial as well as non-financial data [2].

With the rapidly evolving finance environment, uniting the newest machine learning techniques with credit scoring has been a game-changer, revolutionizing the manner in which people and organizations calculate creditworthiness. The intersection of finance and technology has facilitated more accurate and knowledgeable financial decision-making than previously possible. Although conventional credit scoring techniques are still a necessity, they prove limited in offering a complete assessment of the credit risk of an applicant [3]. These traditional models rely mainly on past financial records, neglecting the intricacies of the borrowers' financial conduct and the dynamic nature of contemporary financial systems. Therefore, traditional models tend to disappoint by not representing the true creditworthiness of an individual or a business accurately. This difficulty gave rise to the development of Precision Credit Scoring, a groundbreaking method that takes advantage of complex machine learning processes to turn credit evaluation into an evidence-based science [4]. Precision Credit Scoring basically surpasses typical financial factors in that it taps into a broad range of sources of information. These encompass Internet shopping behaviors, social media behaviors, purchase records, and even biometric signals. By combining this enormous amount of information, this new method builds a complete and elaborate picture of an applicant's financial habits and stability [5].

Credit scoring is an important credit risk assessment tool, employing several techniques to sort applicants into two groups: creditworthy or non-creditworthy. These methods help make credit decisions on loan approvals. Due to this, credit scoring systems are extremely useful for financial institutions and banks, where even small gains can result in quicker decision-making, less risk, and lower credit evaluation costs. With the passage of time, several credit scoring approaches have been devised that can be classified into statistical techniques and artificial intelligence (AI) methods [6]. Statistical techniques follow established assumptions about data, which in some instances are not necessarily similar to actual situations. On the other hand, AI-based methods, especially data mining-based, are able to derive insights directly from data without making any assumptions beforehand, typically providing more accurate and consistent outcomes in credit scoring [7].

Proper evaluation of the creditworthiness of loan applicants and proper risk management are essential to the development of the contemporary financial sector. With the economy's fast growth, conventional statistical credit scoring models have been unable to keep up with the increasing complexity and volume of credit information. Such traditional models tend to make certain statistical assumptions, which do not necessarily apply to large and heterogeneous datasets [8]. By contrast, advances in machine learning and artificial intelligence (AI) have brought more powerful methods of credit scoring. Ensemble learning, evolutionary algorithms, and clustering methods have all been shown great potential in enhancing the accuracy of predictions. In our earlier work, AI-based and machine learning-based models outperformed conventional statistical techniques in building strong credit scoring systems [9, 10].

Imprecise determination of consumer credit risk is crucial to lending organizations because it determines their financial decisions. Credit scoring has become the standard method of enabling financial institutions to measure the likelihood of a credit applicant failing to repay a loan, allowing them to approve or reject credit [14, 15]. An accurate estimate of credit worthiness helps lenders to give out more credit with reduced risks of financial losses. In the last decades, the credit market has developed considerably, with the result of creating sophisticated methods that enable the automatic approval of

credit and keep track of the financial well-being of borrowers. Moreover, because of the high volume of loan requests, modest gains in credit scoring precision can translate into sizeable monetary rewards for lending firms [16].

Machine learning has been applied in credit risk prediction since its inception, using financial information to make an estimate of customers' chances of defaulting on credit cards, other lending products, and loans. Credit risk forecasting is still a foremost challenge for banking institutions, and numerous studies research has focused on enhancing its accuracy [17]. Efficient credit risk prediction solutions can help lenders improve profitability by reducing losses against defaulting clients. Applications for loan and credit cards are major sectors where these tools are heavily used. Institutions that can't properly evaluate credit risk tend to lose heavily financially, thus precise risk evaluation is crucial for their survival. Credit risk forecasting has gained momentum in the last few decades, with default forecasting of credit cards being one of the most serious issues for lenders. This is because default transactions enormously outshine the non-default transactions, thus rendering a natural class imbalance in the credit risk dataset. Earlier work suggests that the imbalances often have adverse effects on the accuracy of machine learning models by perpetuating biases with respect to majority classes and the reduction of default case predictivity [18].

The definition "credit rating" is the determination of a customer's probability of defaulting on debts (Hand & Henley, 1997). The main goal is to place people in two groups: good (those who are financially stable) and bad (those who are most likely to default). Customers who are placed in the financially stable category should be able to honor their repayment undertakings, whereas customers who fall into the high-risk category might not be able to do so [19]. A credit card is among the easiest financial instruments applied in determining the customer's credit rating, which is an indication of the degree of risk they pose. The rating can be measured against a preselected loan approval standard. Because credit evaluation essentially means differentiating between good (low-risk) and bad (high-risk) borrowers, credit scoring typically falls into two broad categories depending on the task at hand and the kind of data utilized (Bijak & Thomas, 2012). The first type pertains to loan applications, in which a prospective applicant's suitability is evaluated according to his or her payment record and money management pattern over time [31]. Banks need to estimate the probability of default within various time horizons, e.g., one month, three months, or six months, in order to remain profitable. The identification of high-risk borrowers allows banks to proactively intervene, e.g., limit credit approvals or adopt risk management measures, to reduce potential losses [20].

With the competitive nature of the current financial environment and the fast-growing lending market, it has become increasingly necessary for there to be proper creditworthiness prediction models in place. Financial institutions, whose main focus is to avert risks while sustaining the stability of their lending portfolio, critically depend on how effective and efficient the prediction methods are [23]. Different machine learning methods, including the logistic regression, Light Gradient Boosting Machine classifier, decision tree classifier, gradient boosting classifier, Extreme gradient boosting classifier and Linear Discriminant analysis, provide distinct strengths and uses in credit risk evaluation. Their application in credit scoring improves predictive power, especially when working with large and intricate datasets [21]. This research also investigates important challenges, such as model interpretability, computational

complexity, and possible limitations, which are important for the practical implementation of these approaches in financial decision-making. Machine learning-based credit risk prediction research is not only of study importance but also have immediate implications for banks, insurance firms, and other sector players. The research outcomes are anticipated to help improve lending choices and enhance risk management practices in the financial industry [22].

Credit risk is a major issue for banks since it has a direct bearing on financial stability. Credit scoring involves a variety of methodologies aimed at helping decision-makers assess loan applicants, enabling financial institutions to reduce losses from non-performing loans. The process identifies whether an applicant is creditworthy or a potential risk [24, 27]. With the growth of banking activities and the accumulation of financial information, credit assessment has moved from manual evaluations to big data-driven and advanced computing technology-based automated credit scoring. For credit risk managers, it is important to differentiate between high-creditworthy borrowers and potential defaulters accurately [29]. Therefore, researchers are still looking for algorithms that improve credit scoring performance. These attempts involve conventional statistical techniques like linear Discriminant analysis and logistic regression, as well as artificial intelligence-based methods like decision trees, artificial neural networks (ANN), Naïve Bayes classifiers, k-nearest neighbors, and support vector machines [26]. While AI-based models have indicated high accuracy in credit scoring, LDA and LR remain in fashion because of their simplicity and ease of use. Yet, these statistical models are criticized for their failure to distinguish between creditworthy and risky applicants, especially in sophisticated financial situations [25].

Credit score classification is a crucial element of the financial sector that helps lenders and institutions determine individuals' and businesses' credit worthiness. Historically, machine learning (ML) methods have dominated the process of creating credit scoring models that offer useful insights for credit risk assessment. With the evolving advances in deep learning (DL), interest has increased in examining its ability to improve credit score classification [28]. DL models, which can automatically learn sophisticated patterns and relationships from data, have much to offer compared to conventional ML methods. By learning complex dependencies in credit data, these models can potentially enhance the accuracy of credit risk predictions and improve financial institution decision-making. This research seeks to provide an extensive comparative evaluation of ML and DL methods in credit score classification [32]. With a heterogeneous dataset for credit histories, we contrast the performance of various ML classifiers, such as RF, decision trees and logistic regression in terms of precision, F1-score, recall, accuracy. Through extensive experimental analysis, we present a detailed comparison of each model's strengths and weaknesses for credit score classification. The results of this research will enable financial institutions to make rational choices when they are choosing classification models that are best suited for their needs. Additionally, we classify these models in terms of the appropriateness of the applications in the financial sector [30].

2. Related Study

In 2024, Emmanuel et al. proposed a stable credit risk prediction model that couples a filter-based feature selection approach based on Information Gain with a stacked ensemble classifier. The model was

evaluated with three benchmark data sets: Australian credit data, German credit data, and Taiwan credit dataset, all from the UCI repository. The stacked classifier was based on three highly effective ensemble algorithms- Extreme Gradient Boosting, Gradient Boosting, Random Forest to achieve optimum predictive performance. To test the model, the authors employed several key metrics including F1-score, Area under the Curve, accuracy. For all datasets, the stacked model outperformed single classifiers consistently, the best performing on the German dataset with 82.80% accuracy, 86.35% F1-score, and 0.944 AUC. Although RF had the best accuracy (87.68%) on the Australian dataset, the stacked classifier had a higher F1-score (84.58%) and AUC (0.934), which reflects more balanced and true predictions. The research concludes that stacking multiple base learners results in better credit risk classification performance, particularly for imbalanced and realistic datasets [1].

Aljadani et al., in their research, introduced an extensive framework integrating machine learning methods and mathematical modeling to improve credit scoring models. They tested several models using three publicly available datasets: South German Credit Dataset, German Credit Dataset and Australian Credit Approval dataset. Algorithms employed are Logistic Regression, Random Forest, K-Nearest Neighbors, Support Vector Machine, Random Forest, Multi-Layer Perceptron (MLP), AdaBoost, XGBoost, LightGBM, and Histogram Gradient Boosting (HGB). The authors proposed a Particle Swarm Optimization-based feature selection approach and employed Adaptive Tree-structured Parzen Estimator (TPE) for optimizing hyperparameters. The main contribution of the work was the interpretation of intricate ML models using the Local Interpretable Model-Agnostic Explanations explainer, thus enhancing transparency in the predictions. From all tested models, RF model attained the accuracy of 88.84% on the Australian dataset, and it performed better with the other models in precision and F1-score. MLP and HGB also performed well on the South German dataset, where MLP attained an accuracy of 77.80%. This paper demonstrates the power of ensemble models in credit scoring and discusses how interpretability by LIME is crucial to real-world financial use cases [2].

Bhilare et al. introduced an integrated credit scoring system using ML methods to improve accuracy and clarity of financial choices. The research applied and contrasted four classification models, namely Support Vector Machine, Naive Bayes, Random Forest, and Decision Tree. Their strategy applied a mixed dataset of 10,000 instances comprising conventional and behavioral financial metrics. Performance was measured by means such as F1-score, recall, precision, accuracy, AUC, MSE, and default rates. From all the models, Random Forest registered the maximum accuracy of 92.5%, then came Decision Tree (91.8%) and SVM (89.2%). RF also reported the lowest default rate (5.2%) and robust performance in precision (0.90) and AUC (0.92), as expected of a strong model. Though SVM proved high in recall and interpretability, Naive Bayes was behind in general performance. The results affirm the dominance of ensemble techniques such as Random Forest in credit risk forecasting and highlight their usefulness in supporting risk management processes in financial institutions [3].

Moral-García and Abellán examined in their work how diversity among base classifiers can be increased to improve credit scoring performance in ensemble approaches. The work introduced a novel approach using the Random Credal Decision Tree (RCDT), which is an extension of the Credal Decision Tree algorithm, where the random hyperparameter value is chosen at training time to induce diversity. RCDT was compared with other classifiers like Logistic Regression (LogR), C4.5, and typical CDTs, using five

ensemble methods: Random Subspace, Bagging, Rotation Forest, DECORATE, AdaBoost. Six credit datasets, namely German, Australian, and UCSD, were experimented on and proved that RCDT performed better than all other models. It recorded the highest average accuracy and AUC on the majority of datasets and ensemble types and performed best on Rotation Forest and Random Subspace ensembles. The authors deduced that RCDT is superior in performance owing to its diversity-causing strategy and is most appropriate for credit scoring purposes in the financial industry [4].

Jin et al. proposed an innovative one or more stage ensemble model coupled with a hybrid genetic algorithm to overcome the limitations of imbalanced datasets in credit score. The model improves forecasting through the use of a Voting-based Instance Hardness Threshold (VIHT) to enhance data sampling quality first. It then utilizes a tailored HGA for feature and classifier selection to ensure only useful data and high-performing classifiers are utilized. Lastly, a stacking ensemble approach is employed to combine predictions for more accuracy. Implemented on three typical imbalanced credit data sets—German, Polish 1, and Polish 2—the model outperformed conventional models and benchmark ensemble strategies on several assessment metrics including Balanced Accuracy, Recall, G-mean and F-score. The findings demonstrate the model's strength and versatility, potentially a useful tool for applications in real-world credit risk evaluation. [5].

João A. Bastos built, in his work, a credit scoring model using Boosted Decision Trees (BDT) and comparing its performance with Multilayer Perceptron (MLP) and Support Vector Machines, applying the same models to the credit datasets German and Australian from the UCI repository. It uses the Area Under the ROC Curve as a parameter for evaluating best performance. BDT had the best accuracy with AUC of 94.0% on the Australian dataset and 81.1% on the German dataset, better than both SVM and MLP. The research also shed light on the most impactful features of credit decision-making. Bastos concluded that BDTs provide a consistent and competitive real-world application for credit scoring because of their robust classification power and interpretability [6].

In Simon Williams' (2021) study, the author compared the performance of ML algorithms in credit card score prediction with a publicly available bank loan data. Models used were Support Vector Machine, Random Forest, Decision Tree, and Logistic Regression. Accuracy and classification report metrics were used to evaluate performance. Out from all models, Random Forest performed the best with 91.32% accuracy and hence is the most accurate model in this research for credit score classification. The worst performance was by the SVM model, especially because it is sensitive to its kernel settings and the structure of the data. Williams concluded that ensemble models such as Random Forest are strong and explainable solutions to real-world credit risk prediction and ought to be used in preference in financial forecasting systems [7].

In their research, Mukhanova et al. (2024) evaluate the application of ML algorithms to predict borrower creditworthiness based on a Kaggle-based credit dataset of 65 features. Compared models are Linear Discriminant Analysis (LDA), Decision Tree, Gradient Boosting, Logistic Regression, XGBoost, and LightGBM. ROC_AUC, F1-score, recall, precision, Accuracy were the evaluation metrics used to compare performance. Among all the models, XGBoost performed the best, with accuracy of 88.4%, outperforming other models on all evaluation metrics. The authors also showed that threshold tuning

could further improve model precision, especially under class-imbalanced scenarios. The research concludes that XGBoost, when paired with appropriate calibration methods, is an extremely effective model for credit scoring tasks [8].

Zou and Gao (2022) proposed a hybrid ensemble model named AugBoost-ELM that augmented Extreme Learning Machine using GB to enhance credit score accuracy. The model was tested on four credit datasets: Taiwan, Japanese, German, Australian and compared against traditional classifiers such as Decision Tree, Logistic Regression, XGBoost, Random Forest, SVM. In the baseline models, Random Forest attained accuracy of 84.35% when tested on Australian Credit dataset. Even though this was marginally less than the proposed AugBoost-ELM model, it also exhibited good performance, reaffirming the dependability of Random Forest in financial forecasting tasks [9].

In their research, R. Shukla, R. Sawant, R. Pawar provides a comparison and analyze of deep learning (DL) and machine learning (ML) methods for credit score classification. The authors compared models like Logistic Regression, Recurrent Neural Network (RNN), Multi-Layer Perceptron (MLP), Convolutional Neural Network (CNN), Random Forest (RF), CNN-RNN hybrid and Decision Tree. Among them, Random Forest provided the maximum accuracy of 90.27%, beating both ML and DL models. MLP and CNN also did great with 87.08% and 87.16% accuracy levels respectively. The research documented the ability of DL in dealing with intricate patterns, whereas RF had a very good balance between recall and precision. RNN, in spite of its time series modeling, had poor defaulting recall. The research offers significant findings for financial institutions that are in need of stable models when it comes to credit risk assessment. This comparative framework assists in choosing appropriate models according to institutional needs such as accuracy, recall, or capability to deal with non-linear data [10].

3. Methodology

3.1 The Proposed Methodology Flowchart for Credit Score Classification:

Fig. 1 illustrates the workflow for classification for credit score using ML techniques. It starts with a dataset that undergoes data preprocessing followed by feature selection. Multiple classification models are then selected for training, including **XGBoost, Logistic Regression, Decision Tree, LightGBM and Random Forest**. The models are evaluated using performance metrics. Among all, **Random Forest is identified as the best-performing model** based on evaluation results. This structured pipeline ensures systematic comparison and selection of the most accurate algorithm for credit scoring tasks.

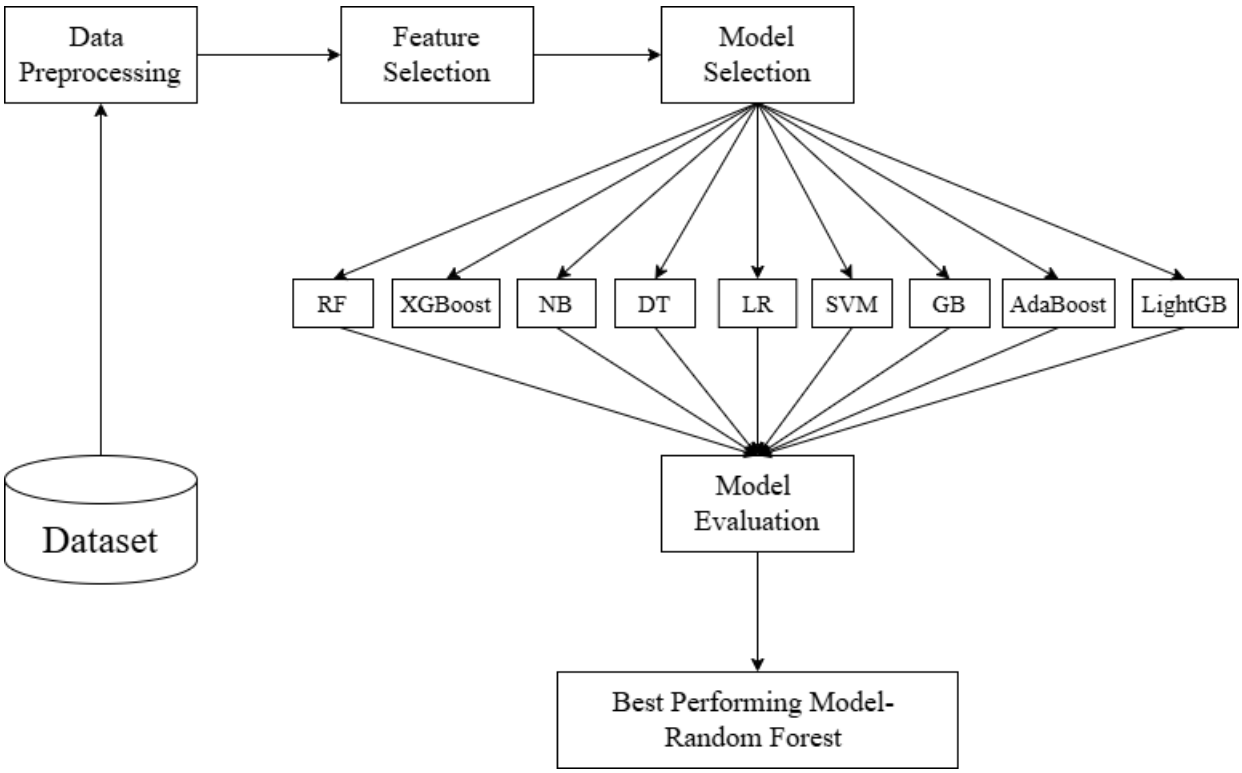

Fig. 1: Flowchart for Credit Score Classification using Machine Learning algorithms.

3.2 Credit Score Classification Kaggle Dataset:

3.2.1 Dataset Description:

The dataset is used in this research study is obtained from **Kaggle** and dataset is intended for **credit score classification**. It consists of approximately **100,000 rows** and **28 attributes**, including both demographic and **financial** details of customers. The dataset includes both **numerical and categorical variables**. The target variable is **Credit_Score**, which is classified into three categories: **Good**, **Standard**, and **Poor**. This dataset provides a comprehensive base for evaluating the performance of various machine learning models in predicting credit risk [11].

3.2.2 Data Preprocessing for Credit Score Classification Dataset:

To ensure the reliability and quality of the model predictions, extensive data preprocessing was performed on the credit score classification dataset. Several attributes initially had **missing values**, which were handled using techniques such as **mean or mode imputation**, depending on the feature type. Certain numerical columns like Age, Outstanding_Debt, and Annual_Income were originally stored as object types and were **converted to numeric formats** after cleansing. Invalid values such as negative ages or placeholder characters were identified and removed. Categorical variables such as Occupation, Credit_Mix, and Payment_Behaviour were **label encoded** based on their cardinality. Outliers in numerical features like Monthly_Inhand_Salary and Credit_Utilization_Ratio were detected and either capped or removed using IQR-based filtering. Text-based features such as

Credit_History_Age were parsed and convert into numerical values (e.g., months). The dataset also checked for **duplicate records** and unnecessary columns like Name, SSN, and ID were dropped to reduce noise. Finally, the features were **scaled** using standardization techniques to ensure uniformity across all models.

Table 1: Credit Score Classification Kaggle Dataset's attributes and its description:

Features	Description	Data-type
ID	Unique hexadecimal identifier for each transaction or entry	Numerical
Customer_Id	Unique identifier assigned to each customer	Numerical
Month	Month of the record/transaction (e.g., January, February)	Categorical
Name	Full name of the customer	Categorical
Age	Age of the customer	Numerical
SSN	Social Security Number (masked; used for unique identification)	Numerical
Occupation	Employment type or job role of the customer.	Categorical
Annual_Income	Total yearly income of the customer	Numerical
Month_Inhand_Salary	Income left in hand after deductions per month.	Numerical
Num_Bank_Accounts	Number of bank accounts held by the customer.	Numerical
Num_Credit_Card	Number of credit cards owned	Numerical
Interest_Rate	Interest rate applied to outstanding credit	Numerical
Num_of_loan	Total number of active loans held by the customer	Numerical
Type_of_loan	Types of loans the customer has.	Categorical
Delay_from_due_date	Number of days payment was delayed beyond the due date	Numerical
Num_of_Delayed_payment	Number of delayed payments made historically	Numerical
Change_Credit_Limit	Change observed in the customer's credit limit over time	Numerical
Num_Credites_Inquiries	Number of inquiries made into the customer's credit report	Numerical
Credit_Mix	Type of credit used (e.g., Good, Standard, Poor mix)	Categorical
Outstanding_Debt	Total unpaid debt amount remaining	Numerical
Credit_Utilization_Ratio	Ratio of used credit to available credit limit	Numerical
Credit_History_Age	Duration of the customer's credit history (e.g., "10 Years and 5 Months")	Numerical
Payment_of_Min_Amount	Whether the customer paid only the minimum amount due (Yes/No)	Categorical
Total_EMI_per_month	Total Equated Monthly Installments paid by customer.	Numerical
Amount_Invested_Monthly	Average monthly investment made by the customer.	Numerical
Payment_Behavior	Spending behavior and payment pattern (e.g., High/Low spent & payment value)	Categorical
Monthly_Balance	Balance amount remaining at the end of the month	Numerical

Credit_Score	Target label indicating creditworthiness (e.g., Good, Standard, Poor).	Categorical

3.2.3 Feature Selection for Credit Score Classification Dataset:

During this process, redundant and irrelevant attributes were removed to improve model performance and efficiency. In particular, the features 'ID', 'Month', 'Name', 'SSN', 'Credit_History_Age', and 'Annual_Income' were eliminated. These variables had high-cardinality identifiers, were uninformative for classification, or included a high percentage of missing or inconsistent values. Removing such features reduces noise and computational complexity while training. The rest of the features still had a considerable predictive capability for credit score classification. Manually, the feature selection was conducted through data exploration and domain applicability. The process was taken to ensure that only the useful attributes fed into the model learning process. Multiple ML models were then train using the cleansed feature set.

3.3 Australian Credit Dataset:

3.3.1 Dataset Description:

The **Australian Credit dataset** is popular benchmark dataset widely used for its evaluating credit scoring models. It contains **690 records** and **14 input features** along with a **binary target variable** that indicates whether a credit application was accepted (1) or rejected (0). The dataset includes a mix of numerical and categorical attributes, capturing key aspects of an applicant's financial and personal background [12].

The features represent information such as the applicant's age, employment status, duration with employers, housing situation, investment history, and banking relationship. Some attributes also reflect monthly financial obligations and savings balance. Despite its compact size, the dataset offers a realistic challenge with its diverse feature types and **presence of missing values**, making it suitable for validating machine learning models under real-world conditions.

The **target variable**, labeled as Class, is binary and helps classify whether an applicant is creditworthy. This dataset provides a strong foundation for implementing and comparing various classification algorithms aimed at credit decision-making.

3.3.2 Data Preprocessing for Australian Credit Dataset:

To prepare the dataset to be used for training and evaluation of the model, a series of preprocessing steps were undertaken. First, the dataset is scanned for the missing or inconsistent values, which were found in numerical and categorical features. Missing values in numerical columns were handled by mean imputation, and missing entries in categorical fields were filled using mode imputation.

The data has a mix of feature types, so categorical variables had to be converted to numerical form. Label encoding was used to do this by assigning integer labels to every category. Continuous variables

like Age, Mean time at addresses, and Monthly housing expense were scaled down to a common scale to ensure better performance for distance-based and gradient-based algorithms.

Further, the dataset was checked for noise and outliers, particularly in financial attributes. Outliers were deleted or rescaled where needed, using statistical cutoffs. Excess identifiers and unclear values were also cleaned for better model interpretability. Finally, the target variable Class was converted into binary (0 = rejection, 1 = approval), and the dataset is divided into training sets and testing sets to compare performance of different classification models under identical conditions.

Table 2: Australian Credit Dataset's attributes and its description:

Feature Name	**Data-Type**	**Description**
Sex	Categorical	Applicant's Gender.
Age	Numerical	Applicant's Age in years.
Mean time at addresses	Numerical	Average duration of residence at current addresses.
Home Status	Categorical	Ownership status of the house (e.g., rented, owned).
Current occupation	Categorical	Job type or occupation of the applicant.
Current job status	Categorical	Employment status (e.g., full-time, part-time).
Mean time with employers	Numerical	Average time spent with different employers.
Other investments	Categorical	Additional assets or investments owned.
Bank account	Categorical	Indicates if the applicant has a bank account.
Time with bank	Numerical	Duration of banking relationship in years.
Liability reference	Categorical	Other financial liabilities the applicant has.
Account reference	Categorical	Reference to other financial accounts.
Monthly housing expense	Numerical	Monthly expenses for housing (rent/mortgage).
Savings account balance	Numerical	Balance in the applicant's savings account.
Class (Reject/Accept)	Categorical	Credit decision: whether the applicant is accepted or rejected.

3.3.3 Feature Selection Australian Credit Dataset:

In order to enhance model performance and simplicity, a subset of features that were relevant from the initial dataset was chosen. The dataset started with 15 attributes, as per Table 2, and had both nominal and continuous variables. Feature selection was based on the scope of relevance of each variable towards

assessing creditworthiness, as well as their possible predictive ability and contribution to interpretability of the model.

The features retained for model training and evaluation are as follows: Other Investments, Time with Bank, Mean Time with Employers, Bank Account, Savings Account Balance, Age, Monthly Housing Expenses, Mean Time at Address, Current Occupation, Current Job Status, and Account Reference. The features were chosen based on a direct or indirect effect on one's financial behavior, stability, and credit risk.

On the other hand, characteristics like Sex, Home Status, and Liability Reference were not included. These were evaluated to be either less pertinent to the credit scoring task, perhaps redundant, or not suitable because of ethical reasons for handling bias and fairness in model prediction. By concentrating on features with high financial and behavioral relevance, the data were filtered to improve training efficiency without sacrificing the most effective predictors for credit classification.

3.4 German Credit Dataset:

3.4.1 Dataset Description:

The **German Credit dataset**, sourced from the UCI ML Repository and originally compiled by Professor Dr. Hans Hofmann, consists of **1,000 records** and **20 input attributes** along with a binary target variable representing credit risk classification. The dataset aims to evaluate whether an individual presents a good or bad credit risk, labeled as 1 for **Good** and 2 for **Bad [13]**.

This dataset includes a mix of 13 categorical attributes and 7 numerical attributes related with a person's financial history, employment status, personal background, and current financial obligations. Key attributes include duration in months, credit amount, employment length, age, and purpose of the loan.

A unique aspect of this dataset is the inclusion of a **cost matrix**, where misclassifying a bad customer as good incurs a higher penalty than the reverse. This makes it particularly suitable for cost-sensitive classification models in the financial domain.

3.4.2 Data Preprocessing for German Credit Dataset:

The German Credit dataset required several preprocessing steps before it could be used for train and evaluating ML models. This dataset contains a combination of **categorical attributes and numerical attributes**, many of which are represented using coded strings (e.g., A11, A12) that needed to be decoded or numerically encoded for modeling purposes. All categorical features were transformed **into numerical format** through techniques such as **label-encoding** or **one-hot encoding**, depending on their cardinality and relevance to the chosen algorithms.

Numerical features such as Duration_in_Month, Credit_Amount, and Age were assessed for **outliers**, and values outside realistic ranges were either corrected or removed to reduce noise. To ensure fair

learning, continuous features were standardized or normalized, especially for algorithms sensitive to feature scales.

The dataset was also checked for **missing or inconsistent values**, although it is generally well-structured and does not contain many null entries. Redundant or non-informative columns were excluded, and the target variable **CreditRisk** was re-encoded as binary for consistency across all models. Lastly, the data was **split into test sets and train sets**, to ensure stratified sampling so that the original class balance was preserved during evaluation.

3.4.3 Feature Selection for German Credit Dataset:

To improve model performance and interpretability, feature selection was performed on the German Credit dataset, which initially had 20 variables. Domain knowledge, correlation analysis, Chi-square testing, and model-based importance (Random Forest) were combined to select the most important predictors. Consequently, 16 attributes were chosen for model training and testing: Age in Years, Duration in Month, , Status of Existing Checking Account, Credit Amount, Present Employment Since, Credit History, Purpose, Personal Status and Sex, Other Debtors/Guarantors, Job, Present Residence Since, Property, Savings Account/Bonds, Installment Rate in % of Disposable Income, Number of Existing Credits at This Bank, and Number of People Being Liable to Provide Maintenance. These characteristics capture important dimensions of a borrower's financial history and credit history. Irrelevant or low-impact characteristics were discarded to enhance training efficiency and prevent model noise.

Table 3: German Credit Dataset's attributes and its description:

Feature Name	Description	Data-Type
Attribute 1	Status of existing checking account	Categorical
Attribute 2	Duration in month	Numerical
Attribute 3	Credit history	Categorical
Attribute 4	Purpose	Categorical
Attribute 5	Credit amount	Numerical
Attribute 6	Savings account/bonds	Categorical
Attribute 7	Present employment since	Categorical
Attribute 8	Installment rate in percentage of disposable income	Numerical
Attribute 9	Personal status and sex	Categorical
Attribute 10	Other debtors/guarantors	Categorical
Attribute 11	Present residence since	Numerical
Attribute 12	Property	Categorical
Attribute 13	Age in years	Numerical
Attribute 14	Other installment plans	Categorical
Attribute 15	Housing	Categorical
Attribute 16	Number of existing credits at this bank	Numerical
Attribute 17	Job	Categorical

Attribute 18	Number of people being liable to provide maintenance	Numerical
Attribute 19	Telephone	Categorical
Attribute 20	Foreign worker	Categorical
Attribute 21	Target Variable	Categorical

3.5 Modeling

The data was split into training and test sets, using 80% of the data for training and the remaining 20% for testing. The models were trained on the training data to learn credit behavior-related patterns and tested using the test set. To determine the performance of various classification methods in estimating credit scores, various machine learning algorithms were used, e.g., XGBoost, Logistic Regression, Decision Tree, Gradient Boosting, Random Forest, LightGB, Adaptive Boosting, Naive Bayes, and Support Vector Machine. These classifier's performances were measured using critical parameters such as F1-score, recall, precision and accuracy to find the most appropriate model for precise credit score classification.

3.5.1 Random Forest:

It is a strong and popular ensemble learning model that forms the backbone for this study. It avoids the issue of overfitting by averaging the predictions of several decision trees, for each learned with the different subsets of data and feature space. This variety among trees strengthens the model's generalization capacity and enhances the accuracy for classification.

The Random Forest algorithm makes its final prediction based on the collective power of an ensemble of n Decision Trees (DTs). Decision Tree is supervised learning model widely applied for both regression and classification problems. In Decision Tree representation, there are three primary node types: root node, which is for starting point of the tree and the first feature split; decision nodes, which are intermediate branching points according to certain feature conditions; and leaf nodes, which give the final output or class label.

In Random Forest, every decision tree in the group independently makes a prediction from a particular input vector X. A majority voting scheme is then used to obtain the overall prediction, which is the class label forecasted by the majority of the decision trees. Mathematically, the RF model is defined as RF = $\{f(X, d_i)\}$, in which i is for the index of decision trees, X is the input data instance, and d_i is the collection of decision functions. The class that is most voted by the ensemble is selected as the model's prediction.

Let us assume we have a data set D containing N samples and M features:

D = $\{(x\_1,\_1),(x\_2,y\_2),...,(x\_N,y\_N)\}$

- x_i is the feature vector of the i-th sample.
- y_i is the target variable (e.g., credit risk label) of the i-th sample.

The objective of building a Decision Tree (T) is to recursively partition dataset into the branches according to certain feature values. At every node in the tree, the algorithm chooses the best feature and associated threshold that most effectively splits the data. This choice is made by optimizing a selected impurity or error measure. For classification tasks, the most frequently used metrics are Gini impurity and entropy. These metrics direct the tree in generating the most informative and efficient splits to enhance predictive accuracy.

For classification using Gini impurity:

$G(D) = 1 - \Sigma(p\_i)\text{^}2$ (1)

Where, p_i is the fraction of samples belonging to class i in node D.

The Objective is to determine the optimal split (feature F and threshold T) that optimizes the weighted sum of Gini impurity in child nodes:

$Gini\ (D,,) = Gini(D) - [p\_left * Gini(D\_left) + p\_right * Gini(D\_right)]$ (2)

Here:
- p_left is for ratio of samples in the left child node.
- p_right is for ratio of samples in the right child node.
- D_left and D_right are representing the datasets in the left and right child nodes, respectively.

Having divided the node at some optimal point, the Decision Tree further divides the resulting child nodes recursively. This continues until some specified stopping condition is satisfied—(e.g., reaching a maximum tree depth or having a minimum number of samples required at a leaf node). In the case of a Random Forest, several decision trees are built on various bootstrapped subsets of the training data, with random subsets of features (feature bagging) at each split. After training all the trees, individual predictions of all trees are combined, and for classification problems, the final response is attained through a majority voting scheme.

3.5.2 Logistic Regression:

Logistic Regression is a basic ML model typically used for binary classification problems, especially with regard to credit risk evaluation. It determines the likelihood of a specific input instance within some category—i.e., "good credit" or "bad credit"—in terms of a sequence input features.

Unlike Linear Regression, which gives continuous outputs, Logistic Regression uses the sigmoid (logistic) function to limit predictions between 0 and 1. This enables the model to treat the result as a probability score. The underlying process is to calculate a linear combination for input variables and apply with the result to the sigmoid function, as shown below:

$$P(y = 1|X) = \frac{1}{1+e^{-(w0+w1x1+w2x2+\cdots+wnxn)}} \quad (3)$$

Where:

- P(y=1|X) is the probability of the positive class (for example, high credit risk),
- $x1$, $x2$,..., xn are the input features (such as income, credit history, loan amount, etc.),
- $w0$ is the intercept or bias term,
- w1, w2,..., wn represent the model's learned weights associated with each feature.

In credit scoring models, Logistic Regression is greatly cherished for its interpretability and simplicity. It enables analysts and banks to see which variable affects the credit decision and aids in support of interpretability and regulatory requirements. Although it does not model complex nonlinear interactions like ensemble methods, it tends to achieve good performance with well-preprocessed data and acts as a good baseline in credit classification problems.

3.5.3 Naive Bayes:

It is a Bayes' Theorem-based probabilistic classification technique commonly applied to applications that demand quick and efficient decision-making. In credit score classification, it offers a simple yet efficient method of predicting whether a person belongs to a high-risk or low-risk credit category.

The model assumes "naive" that all the input features are conditionally independent with respect to the target class, which simplifies the computation. Despite this strong assumption, Naive Bayes is likely to achieve competitive performance, especially on datasets with categorical or text-like features.

Bayes' Theorem is the foundation for the model and this theorem is mathematically expressed as:

$$P(C_k|X) = \frac{P(X|C_k)P(C_k)}{P(X)} \quad (4)$$

- $P(C_k|X)$ represents the posterior probability of class C_k given the feature vector X,
- $P(X|C_k)$ refers to the probability of encountering the attributes under class C_k,
- $P(C_k)$ is the priori probability of class C_k,
- $P(X)$ is the probability of the characteristics (used as a normalizing constant).

In credit risk modeling, Naive Bayes is able to handle the large amount of data rapidly and recognize patterns in financial traits like credit history, income level, or the purpose of a loan. It performs well if the dataset is made up of discrete features or features are mapped appropriately.

3.5.4 Decision Tree:

This Model is a supervised learning model commonly used for both classification and regression problems. Decision Tree works by repeatedly partitioning the dataset into subsets according to that feature maximum information gain or maximum impurity reduction. An internal node of the tree

represents a choice with a feature, branches represent potential outcomes of a choice, and leaf nodes represent the ultimate classification outcome.

The technique of selecting the optimal feature by which to partition the data usually relies on values such as Gini Impurity or Information Gain (Entropy). For classification tasks, the objective is to have pure nodes—where every subset is in a single class as far as possible.

Mathematically, for a dataset D, the Gini Impurity can be calculated as:

$$Gini(D) = 1 - \sum_{i=1}^{n} Pi^2 \qquad (5)$$

- Pi is the probability of class i the dataset.

The algorithm keeps dividing nodes until the stopping criterion is met—like minimum samples at a node, maximum depth, or pure class distribution. While susceptible to overfitting, this can be avoided by using pruning methods or incorporating the model in an ensemble, e.g., Random Forest or Boosting techniques

3.5.5 XGBoost:

XGBoost is a fast, scalable, and flexible machine learning algorithm from the gradient boosting framework. It is now one of the most widely used methods in classification problems, particularly in credit risk modeling, because it can process missing values, avoid overfitting, and provide better predictive performance.

In short, XGBoost creates an ensemble of decision trees one by one. It constructs each new tree to learn from the previous trees' prediction errors. It does so by minimizing a loss function through Gradient descent and each tree try to reduce the residual errors of the existing model.

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + f_t(x_i) \qquad (6)$$

Where:

- $\hat{y}_i^{(t)}$ is representing prediction for instance i at iteration t.
- $f_t(x_i)$ is the output of the newly added tree.

The optimization objective includes both the **loss function** (for example, log loss for classification) and a **regularization term** for control model complexity:

$$Obj = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{t=1}^{T} \Omega(\mathrm{f_t}) \qquad (7)$$

Where:

- l is the loss function(e. g., logistic loss)
- $\Omega(f_t)$ is the complexity of tree f_t

In credit score classification, XGBoost is particularly good at detecting nonlinear relationships and interactions between variables, like income, credit utilization, and payment history. Its resistance to overfitting is due to methods like tree pruning, shrinkage (learning rate), and column sub sampling.

In addition, XGBoost also offers feature importance rankings, allowing analysts to see which financial metrics most affect creditworthiness. This not only makes the model strong but also informative for insights and decision-making in financial uses.

3.5.6 LightGB:

LightGBM model is a light but high-performance Microsoft gradient boosting framework that can deal with large amounts of data with high speed while requiring low memory usage. LightGBM is an enhanced Gradient Boosted Decision Trees (GBDT) variant that is ideally suited for tasks such as classification of credit score, where efficiency and computational expenses are critical considerations.

LightGBM improves upon the conventional boosting algorithms like XGBoost by introducing two key innovations:

Histogram-based decision tree learning: Rather than directly evaluating continuous feature values, LightGBM buckets them into discrete bins. This compresses memory usage and accelerates training without loss of performance.

Leaf-wise tree growing: In contrast to level-wise tree growing employed by most boosting algorithms, LightGBM increases trees by increasing the leaf that provides the best loss reduction. This gives rise to deeper and more precise trees, especially effective in capturing intricate relations in credit data.

The fundamental mathematical concepts underlying LightGBM are the same as the gradient boosting methodology outlined in the XGBoost section above. This involves iteratively updating predictions by minimizing a regularized loss function through additive decision trees. For the mathematical formulation of said process, please look at the equations outlined in the XGBoost model section of this document.

3.5.7 Gradient Boosting:

Gradient Boosting is a ML technique of high performance used both in regression and classification tasks. It is an ensemble family technique where a single prediction is calculated by combining predictions of multiple weak learners—decision trees, generally. In the task of classification of credit score, gradient boosting is particularly appropriate because it has the capacity to detect complex, non-linear relations between a borrower's financial details and his/her creditworthiness.

The core idea of Gradient Boosting is building models in sequence, wherein each new model attempts to cover the errors made by the last one. This is done by optimizing a provided loss function through

gradient descent, wherein each new tree learns to decrease the residual errors of the overall earlier models.

In contrast to the conventional bagging techniques (e.g., Random Forest), which create trees in parallel, Gradient Boosting creates trees individually, and every tree is set to learn from the errors of the previous trees. This creates a robust prediction model that becomes increasingly better at each iteration.

The mathematical basis of the Gradient Boosting algorithm is consistent with the framework already laid out in the XGBoost section, wherein model predictions are progressively improved iteratively through gradient-based optimization.

3.5.8 AdaBoost:

AdaBoost, It requires many weak classifiers—usually decision stumps (one-level decision trees) and aggregates them into a strong predictive model. For credit score classification, AdaBoost assists in improving accuracy by paying greater attention to hard-to-classify borrowers and making progressively better predictions.

The fundamental principle of AdaBoost is to learn a series of models, and each subsequent model gives more emphasis to the examples that were previously misclassified. All instances have equal weights to start with. The weights for misclassified instances are increased at the end of each iteration so that the following model pays even more attention to those tough cases. The overall prediction is the weighted vote of all the individual weak learners.

3.5.9 Support Vector Machine:

Support Vector Machine is a strong supervised machine learning classifier commonly used on classification tasks. SVM is particularly effective in high-dimensional spaces and is well-known for its ability to produce extremely accurate classification boundaries. In the classification of credit scores, SVM can be utilized to classify loan applicants into various credit risk segments based on a set of financial indicators.

The basic concept of SVM is to find the optimal possible hyperplane that can have the maximum margin between two classes. The hyperplane acts as a decision boundary, and the model not only classifies the training set but also generalizes well to new instances. The support vectors are the points on the boundary and play the most significant role in deciding its position. The following support vectors are used for the techniques:

$$w_0^T x + b_0 = 1 or w_0^T x + b_0 = -1 \qquad (8)$$

3.6 Performance Evaluation Measures:

The Performance of the machine learning models was evaluated by different metrics, such as:

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN} \quad (9)$$

$$\text{Precision} = \frac{TP}{TP+FP} \quad (10)$$

$$\text{Precision} = \frac{TP}{TP+FN} \quad (11)$$

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (12)$$

4. Experimental Results and Discussion:

A multiple comparison of models and data sets is, thus, paramount in highlighting the effectiveness of machine learning algorithms and their ability to accurately classify credit scores. A highlight of salient findings appears in this section, with noteworthy implications for the prediction of financial risk and determination of creditworthiness. Random Forest's performance superiority highlights the robustness of ensemble methods in dealing with variable financial attributes. These results continue to validate the incorporation of smart algorithms in current credit assessment systems.

4.1 Algorithm Performance:

In order to improve the reliability and Generalizability of our credit score classification model, we tested it on three different datasets: Credit Score Classification dataset on Kaggle, Australian Credit dataset, and the German Credit dataset. Testing across multiple datasets enabled us to measure model performance across a range of financial profiles and data structures. We trained nine ML models: Decision Tree, GB, Logistic Regression, SVM, XGBoost, LightGBM, Random Forest, AdaBoost, and LightGB on each of the three datasets using identical evaluation metrics, such as recall, precision, F1-score and accuracy. Comparison of results across datasets gave us a better insight into the Generalizability and robustness of each model across different credit environments. The assessment made sure that results were not skewed towards one dataset and could be used in a wider variety of financial risk assessment contexts. All algorithms were tested to provide consistent results.

4.2 Results on Credit Score Classification Kaggle Dataset:

The performance of all ML models was compared to find the most appropriate model for credit score classification. Models were compared using principal metrics: recall, precision, F1-score, and accuracy and on combined dataset analysis. **Random Forest** outperforming the best, with the highest accuracy of **88.57%,** and robust precision (88.55%) and F1-score (88.46%), which reflect its consistent performance to classify credit scores. XGBoost trailed closely with accuracy at 87.92%, presenting competitive values in all performance metrics. Some models like LightGBM (84.56%) and Decision Tree (80.97%) also presented solid performance. The models of Naïve Bayes (66.74%) and Logistic Regression

(67.04%) exhibited lower accuracy and thus seemed less capable of responding to the depth of financial information. These results emphasize the strength of ensemble learning methodologies, especially Random Forest, to provide stable and accurate credit score predictions on varying datasets.

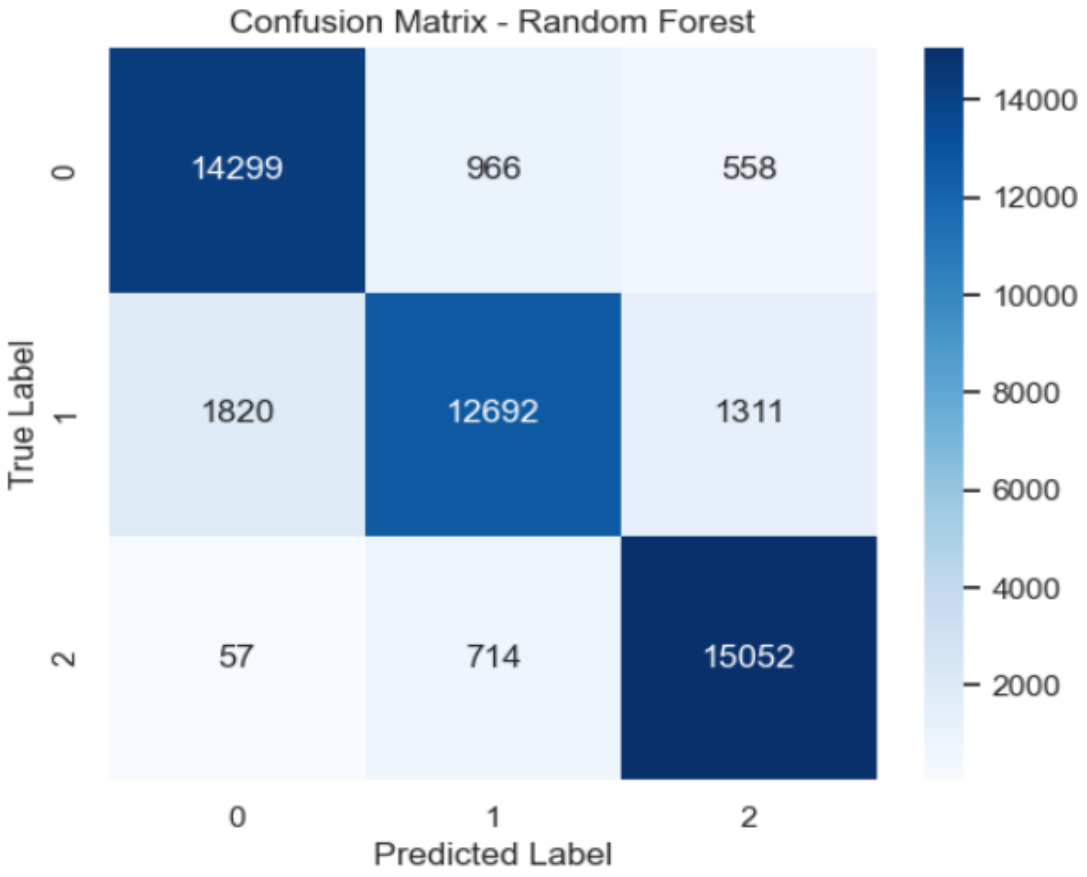

Fig. 2: Random Forest Classifier Confusion Matrix on Credit Score Classification Kaggle Dataset

Table 4: Model Performance on Credit Score Classification Kaggle Dataset:

ML Model	Accuracy	Precision	Recall	F1-score
Random Forest	**88.57%**	**88.55%**	**88.57%**	**88.46%**
XGBoost	87.92%	87.85%	87.92%	87.86%
Decision Tree	80.97%	80.90%	80.97%	80.90%
Logistic Regression	67.04%	66.91%	67.04%	66.68%
SVM	73.10%	73.17%	73.10%	72.71%
LightGB	84.56%	84.47%	84.56%	84.47%
Gradient Boosting	75.10%	75.16%	75.10%	74.88%
AdaBoost	69.35%	69.36%	69.35%	68.99%
Naïve Bayes	66.74%	67.00%	66.74%	66.00%

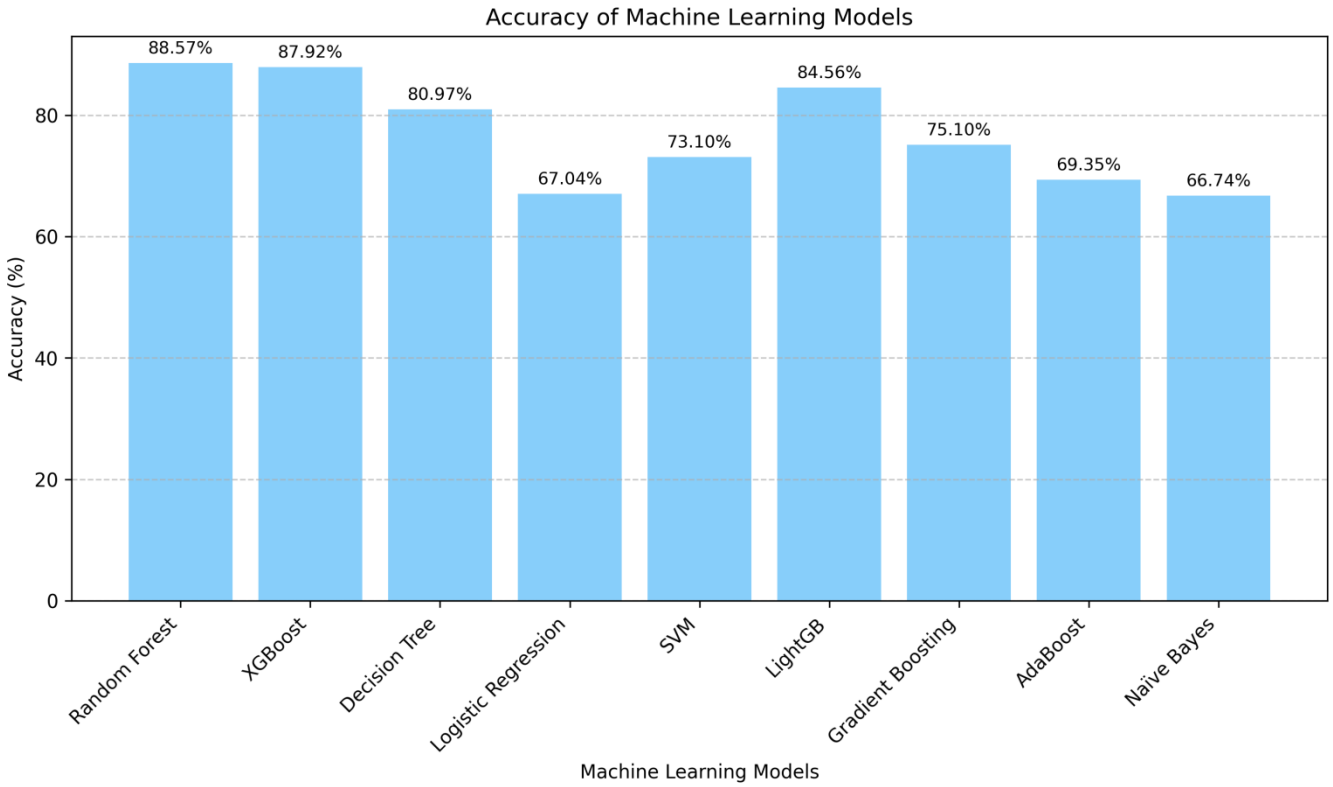

Fig. 3: Model Accuracy on Credit Score Classification Kaggle Dataset

4.3 Results on Australian Credit Dataset:

The performance of all nine ML models was evaluated on the Australian Credit Data dataset to find the best model for credit score classification. Random Forest was the best performing with the highest accuracy of 90.58% and a high F1-score of 91.72%, reflecting consistent classification ability. XGBoost and Gradient Boosting followed closely with the same F1-scores of 89.33%, reflecting competitive precision and recall values. SVM, AdaBoost, and Logistic Regression models also exhibited well-balanced performance with F1-scores of over 86%. LightGBM had consistent results across all the metrics, further establishing its applicability to financial prediction processes. While Decision Tree had the best precision (92.19%), its lower recall had a negative effect on its overall performance.

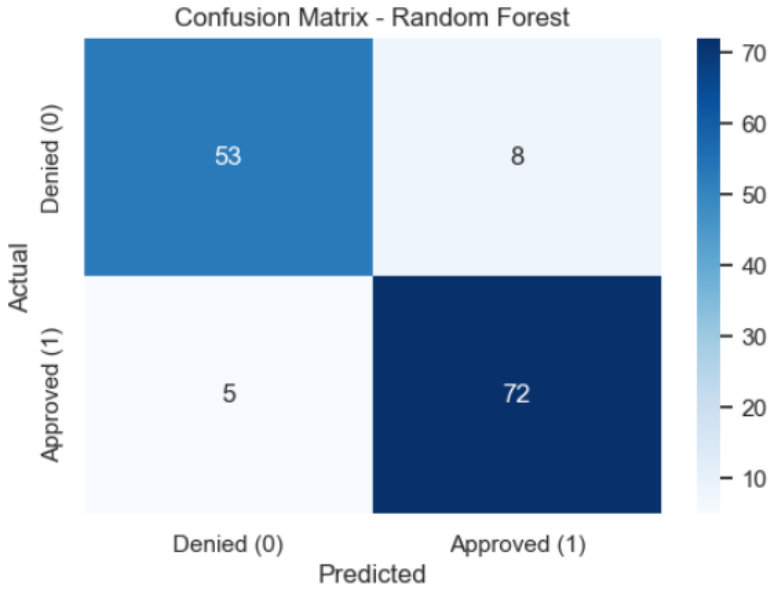

Fig. 4: Random Forest Classifier Confusion Matrix on Australian Credit Dataset

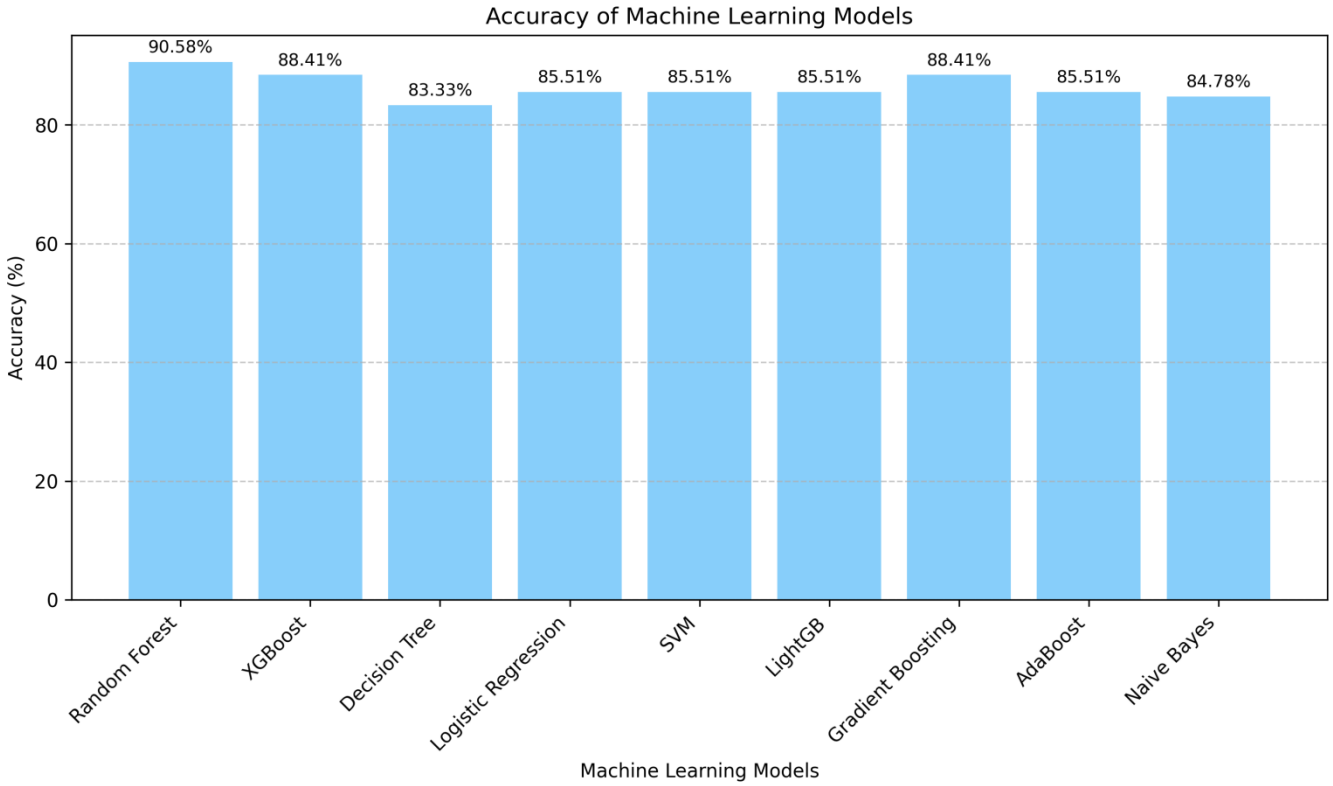

Fig. 5: Model Accuracy on Australian Credit Dataset

Table 5: Model Performance on Australian Credit Dataset:

ML Model	Accuracy	Precision	Recall	F1-score
Random Forest	**90.58%**	90.00%	**93.51%**	**91.72%**
XGBoost	88.41%	91.78%	87.01%	89.33%
Decision Tree	83.33%	92.19%	76.62%	83.69%

Logistic Regression	85.51%	91.30%	81.82%	86.30%
SVM	85.51%	90.14%	83.12%	86.49%
LightGB	85.51%	88.00%	85.71%	86.84%
Gradient Boosting	88.41%	91.78%	87.01%	89.33%
AdaBoost	85.51%	90.14%	83.12%	86.49%
Naive Bayes	84.78%	86.84%	85.71%	86.27%

4.4 Results on German Credit Dataset:

German Credit dataset, multiple machine learning models were compared on core evaluation metrics: recall, F1-score, precision, and accuracy. Out of all models, **Random Forest** showed that best overall accuracy **(88.89%)** and the most robust F1-score (89.22%), indicating its consistency in classifying credit scores with accuracy. XGBoost came second with the F1-score of 88.72%, striking a good balance between precision and recall. LightGBM too performing consistently with an F1-score of 88.06%. Although SVM had a very high recall rate of 93.08%, its precision was slightly less, resulting in an F1-score of 87.05%. Logistic Regression, AdaBoost, and Gradient Boosting performed decently with F1-scores between 85.17% and 86.67%. Decision Tree and Naïve Bayes lagged behind, with lower accuracy and F1-scores, especially Naïve Bayes which had a significant decline in recall (54.62%). These findings bring to fore the potency of ensemble learning models, particularly Random Forest and XGBoost, for stable credit score classification on German Credit Data.

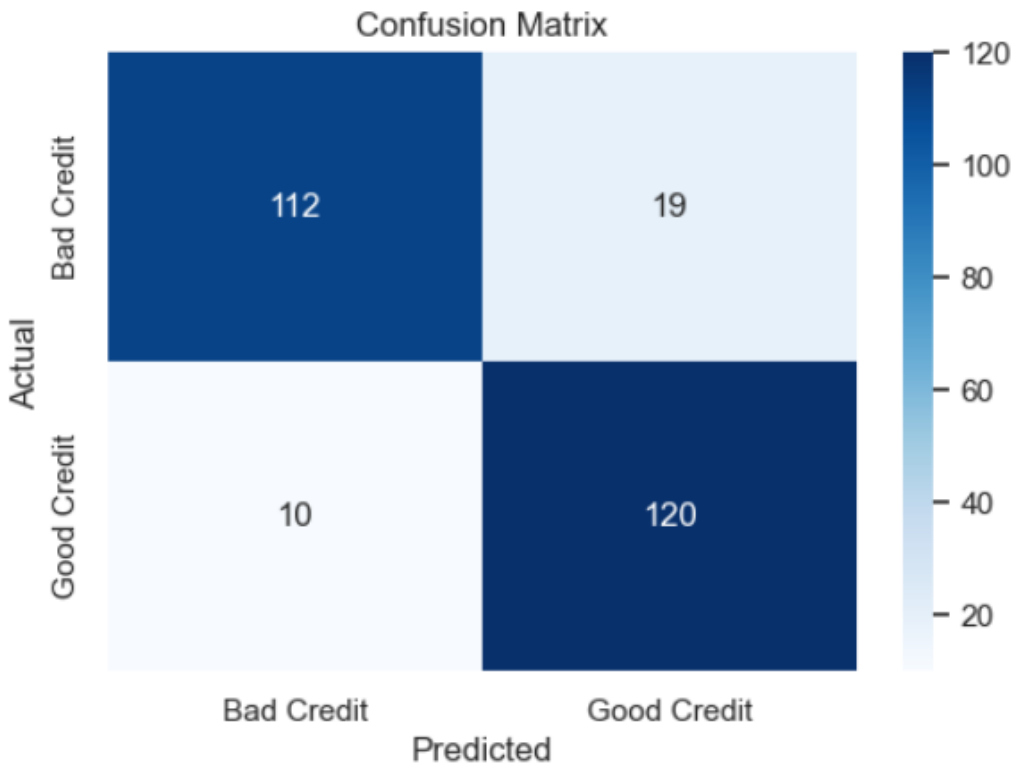

Fig. 6: Random Forest Classifier Confusion Matrix on German Credit Dataset

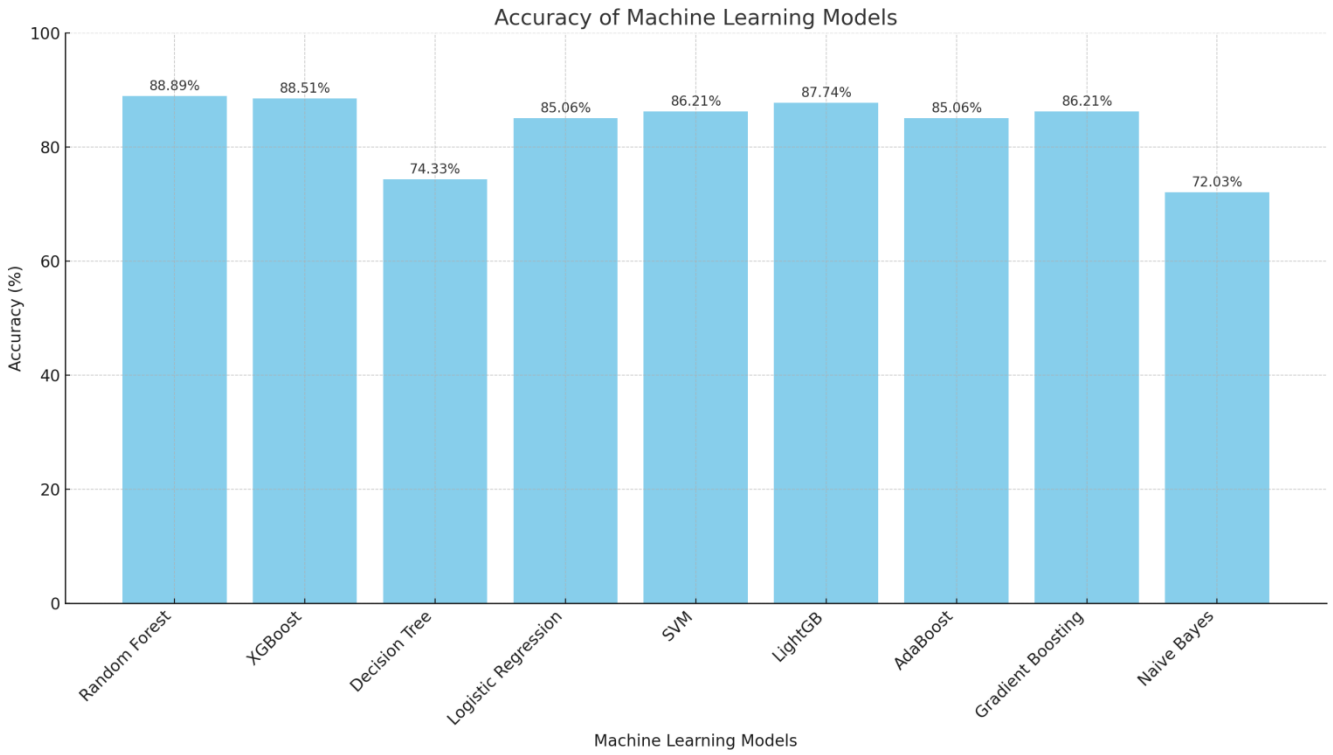

Fig. 7: Model Accuracy on German Credit Dataset

Table 6: Model Performance on German Credit Dataset:

ML Model	Accuracy	Precision	Recall	F1-score
Random Forest	88.89%	86.33%	92.31%	89.22%
XGBoost	88.51%	86.76%	90.77%	88.72%
Decision Tree	74.33%	78.90%	66.15%	71.97%
Logistic Regression	85.06%	83.21%	87.69%	85.39%
SVM	86.21%	81.76%	93.08%	87.05%
LightGB	87.74%	85.51%	90.77%	88.06%
AdaBoost	85.06%	84.21%	86.15%	85.17%
Gradient Boosting	86.21%	83.57%	90.00%	86.67%
Naive Bayes	72.03%	83.53%	54.62%	66.05%

5. Conclusion

This research examined credit score classification with multiple machine learning models on three different datasets: the Kaggle Credit Score Classification dataset, Australian Credit dataset, and German Credit dataset. The accuracy of each model was evaluated based on usual classification metrics: precision, F1-score, accuracy, and recall. Across all three datasets, across all tested models, **Random Forest** was always the best-performing model. When using the Kaggle dataset, it attained **88.57%** accuracy; using the Australian dataset, the best accuracy and F1-score were achieved, at **90.58%** and a

high F1-score of 91.72%, respectively, marking good classification. When using the German dataset, Random Forest achieved **88.89%** accuracy, and it obtained an F1-score of 89.22%. Other ensemble techniques like XGBoost and Gradient Boosting also performed competitively across datasets. For example, XGBoost produced 88.51% accuracy on the German dataset and 87.92% on the Kaggle dataset, coming just behind Random Forest in performance. LightGBM exhibited consistent performance, with F1-scores of more than 84% across all datasets. Classical models like Logistic Regression, SVM, and Naïve Bayes exhibited relatively lower accuracy and F1-scores, especially on the Kaggle dataset, where Logistic Regression achieved only 67.04% accuracy. Comparing with the performance of the overall datasets, Australian Credit dataset had the best classification accuracy, with Random Forest performing at 90.58%, followed by the German dataset at 88.89%, and the Credit Score Classification Kaggle dataset at 88.57%. This indicates that the Australian dataset gave cleaner or more informative features for classification purposes. Overall, **Random Forest** emerged as the most stable and effective credit score classification model on all datasets. The Australian Credit dataset showed the highest performance in all cases, indicating that it was the most appropriate dataset to create stable credit scoring models within this study.

References

1. E. Ileberi, Y. Sun, and Z. Wang, “A machine learning-based credit risk prediction engine system using a stacked classifier and a filter-based feature selection method,” *J. Big Data*, vol. 11, no. 23, pp. 1–14, 2024, doi: 10.1186/s40537-024-00882-0.
2. A. Aljadani, B. Alharthi, M. A. Farsi, H. M. Balaha, M. Badawy, and M. A. Elhosseini, “Mathematical Modeling and Analysis of Credit Scoring Using the LIME Explainer: A Comprehensive Approach,” *Mathematics*, vol. 11, no. 4055, pp. 1–28, Sep. 2023, doi: 10.3390/math11194055.
3. M. Bhilare, V. Wadajkar, and K. Kale, “Empowering Financial Decisions: Precision Credit Scoring through Advanced Machine Learning,” Int. J. Intell. Syst. Appl. Eng., vol. 12, no. 13s, pp. 155–167, Jan. 2024.
4. S. Moral-García and J. Abellán, “Improving the Results in Credit Scoring by Increasing Diversity in Ensembles of Classifiers,” IEEE Access, vol. 11, pp. 58451–58467, 2023, doi: 10.1109/ACCESS.2023.3284137.
5. Y. Jin, W. Zhang, X. Wu, Y. Liu, and Z. Hu, “A Novel Multi-Stage Ensemble Model With a Hybrid Genetic Algorithm for Credit Scoring on Imbalanced Data,” IEEE Access, vol. 9, pp. 143593–143606, 2021, doi: 10.1109/ACCESS.2021.3120086.
6. J. A. Bastos, “Predicting Credit Scores with Boosted Decision Trees,” Forecasting, vol. 4, no. 4, pp. 925–935, 2022, doi: 10.3390/forecast4040050.
7. S. Williams, “Credit Card Score Prediction Using Machine Learning,” *Int. J. Innov. Sci. Res. Technol.*, vol. 6, no. 5, pp. 663–670, May 2021.
8. A. Mukhanova, M. Baitemirov, A. Amirov, B. Tassuov, V. Makhatova, A. Kaipova, U. Makhazhanova, and T. Ospanova, “Forecasting creditworthiness in credit scoring using machine learning methods,” Int. J. Electr. Comput. Eng., vol. 14, no. 5, pp. 5534–5542, Oct. 2024, doi: 10.11591/ijece.v14i5.pp5534-5542.

9. Y. Zou and C. Gao, “Extreme learning machine enhanced gradient boosting for credit scoring,” Algorithms, vol. 15, no. 5, p. 149, Apr. 2022, doi: 10.3390/a15050149.
10. R. Shukla, R. Sawant, and R. Pawar, “A comparative study of deep learning and machine learning techniques in credit score classification,” International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE), vol. 11, no. 7, pp. 10004–10010, Jul. 2023
11. R. Paris, Credit Score Classification, [Dataset], Kaggle, 2022. [Online]. Available: https://www.kaggle.com/datasets/parisrohan/credit-score-classification
12. B. Sigillito, Statlog (Australian Credit Approval) Dataset, [Dataset], UCI Machine Learning Repository, 1994. [Online]. Available: https://archive.ics.uci.edu/dataset/143/statlog+australian+credit+approval
13. B. Sigillito, *Statlog* (German Credit Data) Dataset, [Dataset], UCI Machine Learning Repository, 1994. [Online]. Available: https://archive.ics.uci.edu/dataset/144/statlog+german+credit+data
14. S. Moradi and R. F. Mokhatab, “A dynamic credit risk assessment model with data mining techniques: evidence from Iranian banks,” Financial Innovation, vol. 5, no. 1, p. 15, 2019, doi: 10.1186/s40854-019-0128-8.
15. Y. Song and Y. Peng, “A MCDM-based evaluation approach for imbalanced classification methods in financial risk prediction,” IEEE Access, vol. 7, pp. 84897–84906, 2019, doi: 10.1109/ACCESS.2019.2924751.
16. Z. U. Rehman, N. Muhammad, B. Sarwar, and M. A. Raz, “Impact of risk management strategies on the credit risk faced by commercial banks of Balochistan,” Financial Innovation, vol. 5, no. 1, p. 44, 2019, doi: 10.1186/s40854-019-0157-3.
17. S. Khemakhem and Y. Boujelbene, “Predicting credit risk on the basis of financial and non-financial variables and data mining,” Review of Accounting and Finance, vol. 17, no. 3, pp. 316–340, 2018, doi: 10.1108/RAF-03-2016-0042.
18. V. García, A. I. Marqués, and J. S. Sánchez, “Improving risk predictions by preprocessing imbalanced credit data,” in Neural Information Processing, T. Huang, Z. Zeng, C. Li, and C. Leung, Eds. Berlin, Germany: Springer, 2012, vol. 7664, LNCS, pp. 68–75, doi: 10.1007/978-3-642-34481-7_9.
19. S. Guo, H. He, and X. Huang, “A multi-stage self-adaptive classifier ensemble model with application in credit scoring,” IEEE Access, vol. 7, pp. 78549–78559, 2019, doi: 10.1109/ACCESS.2019.2922676.
20. I. Brown and C. Mues, "An experimental comparison of classification algorithms for imbalanced credit scoring data sets," Expert Systems with Applications, vol. 39, no. 3, pp. 3446–3453, Mar. 2012, doi: 10.1016/j.eswa.2011.09.033.
21. Y. Sayjadah, K. A. Kasmiran, I. A. T. Hashem, and F. Alotaibi, "Credit Card Default Prediction using Machine Learning Techniques," in Proc. 2018 IEEE Conf. on Big Data and Smart Computing (BigComp), Kuala Lumpur, Malaysia, 2018, pp. 1–6, doi: 10.1109/BigComp.2018.00000.
22. N. Torvekar and P. S. Game, "Predictive Analysis of Credit Score for Credit Card Defaulters," Int. J. of Recent Technology and Engineering (IJRTE), vol. 7, no. 5S2, pp. 283–285, Jan. 2019.
23. D. Kikan, S. Singh Jasial, and Y. Singh, "Predictive Analytics Adoption by Banking and Financial Services: The Future Perspective," Int. J. Recent Technol. Eng. (IJRTE), vol. 8, no. 1, pp. 832–834, May 2019.

24. J. He, Y. Zhang, Y. Shi, and G. Huang, "Domain-Driven Classification Based on Multiple Criteria and Multiple Constraint-Level Programming for Intelligent Credit Scoring," IEEE Trans. Knowl. Data Eng., vol. 22, no. 6, pp. 826–838, Jun. 2010, doi: 10.1109/TKDE.2010.43.
25. M. M. M. Megdad and S. S. Abu-Naser, "Credit Score Classification Using Machine Learning," *International Journal of Academic Information Systems Research (IJAISR)*, vol. 8, no. 5, pp. 1–10, May 2024.
26. F. Louzada, A. Ara, and G. B. Fernandes, "Classification methods applied to credit scoring: Systematic review and overall comparison," *Surveys in Operations Research and Management Science*, vol. 21, no. 2, pp. 117–134, 2016, doi: 10.1016/j.sorms.2016.10.001.
27. D. Tripathi, D. R. Edla, A. Bablani, A. K. Shukla, and B. R. Reddy, "Experimental analysis of machine learning methods for credit score classification," Progress in Artificial Intelligence, vol. 10, no. 4, pp. 555–573, 2021, doi: 10.1007/s13748-021-00238-2.
28. Y. Zhong and H. Wang, "Internet Financial Credit Scoring Models Based on Deep Forest and Resampling Methods," IEEE Access, vol. 11, pp. 8689–8703, 2023, doi: 10.1109/ACCESS.2023.3239889.
29. E. Ileberi and Y. Sun, "Advancing Model Performance With ADASYN and Recurrent Feature Elimination and Cross-Validation in Machine Learning-Assisted Credit Card Fraud Detection: A Comparative Analysis," IEEE Access, vol. 12, pp. 133315–133327, Sept. 2024, doi: 10.1109/ACCESS.2024.3457922.
30. G. Teles, J. J. P. C. Rodrigues, K. Saleem, and S. A. Kozlov, "Classification Methods Applied to Credit Scoring With Collateral," IEEE Systems Journal, early access, 2019, doi: 10.1109/JSYST.2019.2937552.
31. M. Malekipirbazari and V. Aksakalli, "Risk Assessment in Social Lending via Random Forests," Expert Systems with Applications, vol. 42, no. 10, pp. 4621–4631, 2015, doi: 10.1016/j.eswa.2015.02.001.
32. Z. Hassani, M. A. Meybodi, and V. Hajihashemi, "Credit Risk Assessment Using Learning Algorithms for Feature Selection," *Fuzzy Information and Engineering*, vol. 12, no. 4, pp. 529–544, 2020, doi: 10.1080/16168658.2021.1925021.