A FIELD PROJECT REPORT

on

**“CREDIT RISK ANALYSIS”**

**Submitted**

by

|  |  |
| --- | --- |
| 221FA04075  P.Praharshitha | 221FA04123  V.Vamsi |
| 221FA04152 221FA04215  P.Pujitha T.Mohan | |
|  | |

**Under the guidance of**

*Dr. Vinoj J*

Assistant Professor

Dept of CSE



**DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING**

**VIGNAN'S FOUNDATION FOR SCIENCE, TECHNOLOGY AND RESEARCH Deemed to be UNIVERSITY**

**Vadlamudi, Guntur.**

**ANDHRA PRADESH, INDIA, PIN-522213.**



**CERTIFICATE**

This is to certify that the Field Project entitled **“CREDIT RISK ANALYSIS”** that is being submitted by 221FA04075(P.Praharshitha), 221FA04123(V.Vamsi), 221FA04152(P.Pujitha),221FA04215(T.Mohan)for partial fulfilment of Field Project is a bonafide work carried out under the supervision of *Dr. Vinoj J ,* Assistant Professor ,Dept of CSE

|  |  |  |
| --- | --- | --- |
| Guide name& Signature |  | Dr.K.V. Krishna Kishore |
| Assistant/Associate/Professor, CSE | HOD,CSE | Dean, SoCI |



**DECLARATION**

We hereby declare that the Field Project entitled **“CREDIT RISK ANALYSIS”** is being submitted by 221FA04075 (P.Praharshitha), 221FA04123 (V.Vamsi), 221FA04152 (P.Pujitha) and 221FA04215 (T.Mohan)in partial fulfilment of Field Project course work. This is our original work, and this project has not formed the basis for the award of any degree. We have worked under the supervision of Dr. Vinoj J, Assistant Professor, Department of CSE.

By

**221FA04075 (P.Praharshitha),**

**221FA04123(V.Vamsi),**

**221FA04152(P.Pujitha),**

**221FA04215(T.Mohan)**

Date:

## ABSTRACT

Credit risk management plays a pivotal role in the financial sector, directly impacting lending decisions and the overall economic ecosystem. This study leverages machine learning algorithms to enhance credit risk prediction using a comprehensive dataset of credit card applications obtained from Kaggle. Several classification models, including Logistic Regression, Support Vector Machines, Decision Trees, Random Forests, K-Nearest Neighbors, Gradient Boosting, Neural Networks, and AdaBoost, were employed to assess the likelihood of loan defaults. The models were systematically trained and validated, with Gradient Boosting demonstrating strong performance in terms of accuracy and predictive capability. Key features, such as credit history, age, and employment status, were identified as significant contributors to predicting credit risk. By analyzing model performance across metrics like accuracy, precision, recall, and F1-score, this research highlights the advantages of integrating advanced machine learning techniques into credit risk analysis. The findings offer valuable insights for financial institutions, enabling more informed decision-making and improved risk management strategies.

**Keywords**: Credit Risk Management, Machine Learning Algorithms, Risk Prediction, Feature Analysis.

**TABLE OF CONTENTS**

**1.Introduction**

**2.Literature Survey**

**2.1 Literature review**

**2.2Motivation**

**3.Proposed System**

**3.1 Input Dataset**

**3.1.1 Detailed Features of the Dataset**

**3.2 Data Preprocessing**

**3.3 Model Building**

**3.4 Methodology of the System**

**3.5 Model Evaluation**

**3.6 Constraints**

**3.7 Cost and Sustainability Impact**

**4. Implementation**

**4.1 Environment Setup**

**4.2 Sample Implementation of models**

**5. Experimentation and Result Analysis**

**6.Conclusion**

**7. References**

**LIST OF FIGURES**

|  |
| --- |
| Figure 1. Architecture of the proposed system |
| Figure 2. Various features in the dataset |
| Figure 3. SVM Confusion Matrix |
| Figure 4. KNN Confusion Matrix |
| Figure 5. Decision Tree Confusion Matrix |
| Figure 6. Extra trees Confusion Matrix |
| Figure 7.Logistic Regression Confusion Matrix  Figure 8.Random Forest Confusion Matrix  Figure 10. Gradient Boosting Confusion Matrix  Figure 11.Neural Network Confusion Matrix  Figure 12.Adaboost Confusion Matrix  Figure 13.Logistic Regression(precision, recall, f1-score)  Figure 14.Random Forest(precision, recall, f1-score)  Figure16. Neural Network()  Figure 17. Gradient Boosting() |
| Figure 18: SVM (precision, recall, f1-score) |
| Figure 19: Decision Tree (precision, recall, f1-score) |
| Figure 20: KNN (precision, recall, f1-score) |
| Figure 21: Extra trees(precision, recall, f1-score) |

# CHAPTER-1

# INTRODUCTION

### INTRODUCTION

In the contemporary financial environment, the importance of precise credit risk evaluation cannot be overstated, especially for lending institutions. As credit transactions surge and borrower behaviors evolve, lenders encounter substantial hurdles in accurately assessing creditworthiness. Conventional techniques for evaluating credit risk often fail to account for the complexity and diversity of today's financial and demographic information. This has intensified the demand for advanced methodologies that can adeptly navigate the nuances of extensive datasets and yield more trustworthy predictions.

This study seeks to tackle these issues by harnessing machine learning algorithms to refine credit risk assessment processes. Utilizing a comprehensive dataset sourced from Kaggle, containing extensive records of credit card applications, we examine a variety of machine learning models, including Logistic Regression, Support Vector Machines, Decision Trees, Random Forests, K-Nearest Neighbors, Gradient Boosting, and Neural Networks. These models are meticulously trained and validated to create robust binary classifiers that predict the likelihood of applicant default.

Through the application of sophisticated analytical methods, this research assesses the performance of each model based on key metrics such as accuracy, precision, recall, and F1-score, while also pinpointing critical features that significantly impact prediction outcomes. The results illustrate the transformative potential of machine learning in financial risk management, providing institutions with data-driven insights that can enhance their credit risk assessment techniques, ultimately fostering a more stable and resilient lending landscape.

# CHAPTER-2 LITERATURE SURVEY

## LITERATURE SURVEY

#### 2.1 Literature review

The application of machine learning and statistical models for credit risk prediction has gained significant traction in recent years. Perera and Premaratne (2016) applied Artificial Neural Networks (ANN) to forecast payment behaviors of leasing customers, highlighting the model's ability to capture non-linear patterns in financial data. Similarly, Adewusi et al. (2016) demonstrated the effectiveness of ANN in predicting loan recovery, emphasizing its potential in financial applications where traditional models may underperform.

Marqués, García, and Sánchez (2012) explored the performance of base classifiers within ensemble methods in credit scoring, noting that ensemble models generally outperform individual classifiers due to their ability to aggregate diverse predictions. This aligns with findings from Aslam et al. (2019), who conducted empirical studies on loan default prediction, showcasing the robustness of ensemble methods for enhancing predictive accuracy.

Atiya (2001) provided an extensive survey on using neural networks for bankruptcy prediction, offering both theoretical insights and new results that underscore the relevance of ANN in credit risk management. Additionally, Pandit (2016) applied data mining techniques on a loan approval dataset to predict defaulters, finding that machine learning approaches improve prediction accuracy compared to traditional statistical methods.

Other studies have examined the effectiveness of various machine learning techniques in predicting loan defaults. For instance, Nehrebecka (2018) compared logistic regression and Support Vector Machines (SVM), concluding that SVM can outperform logistic regression in certain credit scoring tasks. Vimala and Sharmili (2018) further supported this by using both Naive Bayes and SVM, finding SVM to be superior in handling large datasets and providing more accurate predictions.

Baesens, Roesch, and Scheule (2016) discussed a broad range of credit risk analytics techniques, emphasizing their applications in real-world financial scenarios. Byanjankar et al. (2015) employed neural networks in peer-to-peer lending, finding that ANN models were particularly effective in evaluating the creditworthiness of borrowers in online lending environments.

Lastly, Angelini, di Tollo, and Roli (2008) introduced a neural network model for credit risk evaluation, demonstrating its ability to handle complex, nonlinear relationships inherent in credit data. Their work highlighted the adaptability of machine learning models like ANN to financial risk assessment.

#### 2.2 Motivation

In an increasingly digitized world, the financial sector faces unprecedented volumes of data and complexity in credit risk evaluation. As lending institutions expand their services, they must contend with the challenge of accurately predicting credit risk for diverse customer profiles. Traditionally, credit risk assessment relied heavily on static rules and human judgment, which often fell short in recognizing the intricate patterns of financial behavior that lead to defaults. This has resulted in inefficiencies, such as higher default rates and misallocation of credit, both of which undermine the stability of financial systems.

The motivation behind this research lies in addressing the limitations of traditional credit risk evaluation methods by leveraging machine learning algorithms. The ability of these models to process large datasets and uncover hidden correlations offers significant potential for improving prediction accuracy. In particular, we aim to provide financial institutions with more reliable tools for assessing creditworthiness, helping them reduce defaults and make more informed lending decisions. As economic stability depends on effective credit risk management, advancing these techniques becomes critical to sustaining growth and ensuring the robustness of financial markets.

By integrating various machine learning models into credit risk prediction, this study seeks to enhance the accuracy and efficiency of the assessment process, delivering results that are not only data-driven but also adaptable to the evolving landscape of credit lending. This motivation aligns with the broader goal of optimizing risk management strategies, ultimately contributing to more secure and resilient financial systems.

# CHAPTER-3 PROPOSED SYSTEM

### 3. PROPOSED SYSTEM

The proposed system aims to revolutionize credit risk assessment by integrating advanced machine learning techniques into the evaluation process. The system is designed to analyze a comprehensive dataset of credit card applications, leveraging both financial and demographic information to predict whether an applicant is likely to default on their loan or credit card payments. The system consists of several core components that work together to build, train, and validate predictive models, ensuring robust and accurate credit risk predictions.

**Data Collection and Preprocessing**:  
The system uses a rich dataset from Kaggle, which includes various attributes such as credit history, employment status, income level, and personal demographic information. The initial step involves cleaning and preprocessing the data to handle missing values, outliers, and categorical variables. Feature scaling and normalization techniques are also applied to ensure that the data is suitable for the machine learning models.

**Model Selection and Implementation**:  
The system employs a diverse range of machine learning algorithms to evaluate credit risk. The selected models include:

* 1. **Logistic Regression**: A statistical method for binary classification, used as a baseline model to predict defaults.
  2. **Support Vector Machine (SVM)**: A robust classifier that finds the optimal boundary between classes in high-dimensional spaces.
  3. **Decision Tree**: A tree-based model that splits data based on the most significant features to classify loan applicants.
  4. **Random Forest**: An ensemble model that aggregates multiple decision trees to improve predictive accuracy and reduce overfitting.
  5. **K-Nearest Neighbors (KNN)**: A distance-based algorithm that classifies applicants by comparing them with the most similar cases in the dataset.
  6. **Gradient Boosting**: A powerful ensemble technique that builds weak learners sequentially to correct errors, often resulting in high accuracy.
  7. **Neural Network (MLP)**: A multi-layer perceptron that captures complex patterns in the data through interconnected neurons.
  8. **AdaBoost**: An adaptive boosting model that combines multiple weak classifiers to create a strong learner, improving classification performance.

**Model Training and Evaluation**:  
Each model is trained on the preprocessed dataset and fine-tuned through hyperparameter optimization to achieve optimal performance. The system evaluates the models based on several key performance metrics, including accuracy, precision, recall, and F1-score. These metrics provide a comprehensive understanding of each model’s ability to correctly classify applicants into low-risk or high-risk categories.

**Feature Importance and Analysis**:  
 A critical aspect of the system is identifying which features (such as credit history, age, income, or employment status) most strongly influence credit risk predictions. This analysis helps financial institutions understand the underlying factors contributing to loan defaults, allowing for more targeted risk mitigation strategies.

**Model Deployment**:  
Once the best-performing model is identified (based on accuracy and other evaluation metrics), the system can be deployed in real-world applications for live credit risk prediction. Financial institutions can use the system to automate credit approval processes and make more data-driven decisions, ultimately reducing the risk of defaults and improving lending efficiency.

#### 3.1 Input Dataset

The input data for our credit risk analysis consists of two comprehensive datasets from Kaggle: **application\_record.csv** and **credit\_record.csv**. These datasets provide detailed insights into both the demographic and financial profiles of applicants, forming the foundation for our machine learning models designed to predict credit risk.

The **application\_record.csv** dataset contains **438,557 rows** and **18 columns** of demographic and financial information related to credit card applicants. Each row represents a unique applicant, identified by a specific **ID**. This dataset includes critical features such as **gender**, indicating whether the applicant is male or female, and **car ownership**, a binary feature showing whether the applicant owns a vehicle. Additionally, it includes information on **real estate ownership**, revealing whether the applicant owns property, and **income type**, categorizing applicants based on their source of income, such as working, commercial associate, or pensioner.

Other important variables include **education level**, which reflects the applicant’s highest degree of formal education, and **marital status**, providing information about whether the applicant is single, married, or in another marital category. These features contribute to understanding the socioeconomic status of the applicants. Moreover, the dataset records **age**, which is vital for assessing long-term credit risk, and **employment status**, which offers insight into job stability, a key determinant of an individual's financial capability.

The second dataset, **credit\_record.csv**, complements the demographic data by offering a historical perspective on the applicants' credit activities. It includes **1,048,575 rows** and **3 columns**, tracking the monthly credit status of each applicant, such as whether payments were delayed, whether an individual is overdue, or if the account is in good standing. This dataset helps us understand the repayment behavior of applicants, an essential factor in predicting future defaults or financial risks.

Together, these datasets provide a rich source of data for developing machine learning models that can assess credit risk more effectively, helping financial institutions make better lending decisions.

#### 3.1.1 Detailed Features of the Dataset

The dataset used in this study contains a wide range of features that capture both demographic and financial characteristics of credit card applicants, along with historical credit behavior. These features are critical for building predictive models that can assess credit risk. Below is a detailed description of the key features in the dataset:

**1. ID**

A unique identifier assigned to each applicant, used to merge data between the application\_record.csv and credit\_record.csv datasets.

**2. Gender**

Specifies the gender of the applicant. This binary feature is used to assess any potential gender-related patterns in credit risk.

**3. Car Ownership**

A binary feature indicating whether the applicant owns a car. This can be linked to financial stability and creditworthiness.

**4. Real Estate Ownership**

Indicates if the applicant owns property, which can serve as collateral and reflect financial security.

**5. Number of Children**

This feature captures the number of dependent children an applicant has, which may affect their disposable income and ability to repay loans.

**6. Income Type**

Describes the type of income the applicant earns, categorized as working, commercial associate, pensioner, or others. This feature is crucial in understanding the financial background of the applicant.

**7. Income Amount**

The total annual income of the applicant, which is a key determinant of the applicant’s ability to meet repayment obligations.

**8. Education Level**

Captures the applicant’s educational qualifications (e.g., secondary, higher education). This feature is often correlated with income potential and financial responsibility.

**9. Marital Status**

Specifies the marital status of the applicant (e.g., single, married). This feature helps to understand the financial obligations and lifestyle of the applicant.

**10. Housing Type**

This feature indicates the type of housing the applicant resides in, such as rented apartment, house, or office apartment, which can reflect their financial status.

**11. Age**

The age of the applicant, derived from the number of days since birth. Age is a critical factor in credit risk assessment, as younger or older applicants may have different credit behaviors.

**12. Employment Length**

The number of days the applicant has been employed. Longer employment histories often correlate with financial stability and a lower likelihood of default.

**13. Mobile Phone Ownership**

A binary feature that indicates whether the applicant owns a mobile phone. While this may seem basic, it can reflect the level of connectedness and access to services.

**14. Work Phone Ownership**

Indicates if the applicant has a work phone, which may provide insight into their employment status and job stability.

**15. Home Phone Ownership**

Specifies whether the applicant has a home phone, which might reflect their residential stability.

**16. Email Ownership**

A binary feature indicating whether the applicant has an email address, which can be indicative of digital accessibility and modern financial behavior.

**17. Occupation Type**

Captures the applicant’s occupation, which may directly influence their income level and job security.

**18. Family Member Count**

The total number of family members supported by the applicant. Larger families may indicate higher financial obligations, affecting disposable income.

**19. Credit Status (Credit History)**

#### From the credit\_record.csv dataset, this feature tracks the monthly credit status of applicants, such as whether they were overdue, the number of days past due, or if payments were made on time. It is a key determinant in predicting future credit risk. 3.2 Data Pre-processing

In any machine learning project, data pre-processing plays a critical role in ensuring model accuracy and reliability. For this credit risk prediction system, the dataset from Kaggle undergoes a rigorous pre-processing stage to handle inconsistencies and prepare the data for model training. This process involves several key steps, including handling missing values, where records with incomplete information are either imputed or removed, depending on the significance of the missing data. Categorical variables, such as gender, education, and employment status, are transformed into numerical representations using techniques like one-hot encoding. Furthermore, numerical features are scaled and normalized to ensure that all variables contribute equally to model performance, especially in distance-based models like KNN. Outliers are also detected and either capped or removed to minimize their impact on model predictions. These pre-processing steps are essential for ensuring that the models receive clean, well-structured data, thereby improving their ability to accurately predict credit risk.

**preprocessing techniques applied:**

Preprocessing is a critical step in machine learning projects, especially when working with raw data. For our credit risk prediction project, we implemented various preprocessing techniques to clean, transform, and prepare the data for model training. These steps are essential to ensure that the data is suitable for analysis and improves the performance of machine learning models. Below are the key preprocessing techniques applied in the project:

· **Handling Missing Values**:

* · **Imputation**: Missing values in numerical features (e.g., income, age) were imputed using the mean or median, depending on the distribution of the feature. For categorical variables (e.g., education level, marital status), the most frequent category was used for imputation.
* **Removal**: Records with excessive missing values (e.g., more than 20% of features missing) were removed from the dataset to maintain data integrity.

· **Outlier Detection and Treatment**:

* · **Statistical Methods**: Outliers were detected using the Interquartile Range (IQR) method. Values lying beyond 1.5 times the IQR above the third quartile or below the first quartile were considered outliers.
* **Capping**: Outliers were capped at the 95th percentile to mitigate their influence on model training without discarding valuable data.

· **Feature Encoding**:

* · **One-Hot Encoding**: Categorical features, such as gender and employment type, were converted into binary (0/1) variables using one-hot encoding. This allows machine learning models to interpret these features without assuming any ordinal relationship.
* **Label Encoding**: For ordinal categorical features (e.g., education level), label encoding was applied to assign integer values based on the ranking of categories.

· **Feature Scaling**:

* · **Standardization**: Numerical features were standardized to have a mean of 0 and a standard deviation of 1, particularly beneficial for models sensitive to feature scales, such as Logistic Regression and SVM.
* **Normalization**: Min-max normalization was applied to scale features to a range between 0 and 1, which is particularly useful for distance-based algorithms like KNN.

· **Feature Selection**:

* · **Correlation Analysis**: Features were evaluated for correlation with the target variable (default/no default) to identify significant predictors. Features with a low correlation (below a specified threshold) were considered for removal.
* **Recursive Feature Elimination (RFE)**: RFE was applied to select the most important features based on model performance, enhancing model interpretability and reducing overfitting.

· **Data Splitting**:

* · The processed dataset was divided into training and testing sets, typically following an 80-20 split. This ensures that models are trained on a substantial amount of data while retaining a separate set for unbiased evaluation.

#### 3.3 Model Building

Model building is a crucial step in the machine learning pipeline, where we apply various algorithms to predict credit risk based on the preprocessed data. The primary objective of this stage is to train models that can classify applicants into low-risk and high-risk categories. In our project, several machine learning algorithms were employed to determine which model yields the highest predictive performance.  
**Model Selection**:  
Based on the objectives of the study, seven machine learning models were selected for training:

* 1. **Logistic Regression**
  2. **Support Vector Machine (SVM)**
  3. **Decision Tree**
  4. **Random Forest**
  5. **K-Nearest Neighbors (KNN)**
  6. **Gradient Boosting**
  7. **Neural Network (MLP)**
  8. **AdaBoost**

**Training Phase**:  
Each model was trained using the training dataset (80% of the processed data). The training involved:

* 1. **Hyperparameter Tuning**: For certain models, such as SVM and Random Forest, hyperparameters were optimized using techniques like Grid Search or Random Search. This process identifies the best parameters that enhance model performance.
  2. **Cross-Validation**: K-Fold Cross-Validation (e.g., 5 or 10 folds) was employed to ensure that the model's performance is robust and generalizes well to unseen data. Each model was trained multiple times on different subsets of the training data.

**Model Evaluation**:  
After training, the models were evaluated on the testing dataset (20% of the processed data). The following performance metrics were computed:

* 1. **Accuracy**: The ratio of correctly predicted instances to the total instances.
  2. **Precision**: The ratio of true positive predictions to the total predicted positives, indicating the model's ability to minimize false positives.
  3. **Recall**: The ratio of true positive predictions to the total actual positives, reflecting the model's capability to identify actual defaults.
  4. **F1-Score**: The harmonic mean of precision and recall, providing a balanced measure of model performance.
  5. **ROC-AUC Score**: The Area Under the Receiver Operating Characteristic Curve (AUC-ROC) was calculated to assess the model's ability to distinguish between the positive and negative classes.

**Feature Importance Analysis**:  
For tree-based models such as Random Forest and Gradient Boosting, feature importance scores were extracted to identify which features contributed most significantly to the model's predictions. This analysis helps in understanding the underlying factors influencing credit risk and aids in making informed business decisions.

**Final Model Selection**:  
Based on the evaluation metrics, the model demonstrating the highest accuracy and F1-Score was selected for deployment. In this project, Gradient Boosting emerged as the best-performing model with an accuracy of 0.97, indicating its effectiveness in handling complex patterns in the data.

**Model Interpretation**:  
The selected model was analyzed for interpretability, ensuring that stakeholders can understand the decision-making process behind the predictions. Techniques such as SHAP (SHapley Additive exPlanations) values or LIME (Local Interpretable Model-agnostic Explanations) were employed to provide insights into how individual features influenced predictions.

**Deployment Readiness**:  
Finally, the trained model was prepared for deployment in a real-world scenario, including the creation of an API or user interface for integrating the model into the existing credit evaluation systems of financial institutions.

We used an 80:20 train-test split with 10-fold cross-validation to evaluate model consistency and reduce overfitting. Hyperparameter tuning, using grid and random search, helped optimize the models. Performance was assessed using accuracy, precision, recall, F1-score, and ROC-AUC.

**Gradient Boosting** emerged as the best model, achieving an accuracy of **0.90** on the validation set and **0.8559** on the test set, effectively handling the dataset's complexity and class imbalance.

#### 3.4 Methodology of the system

The methodology employed in this study begins with a comprehensive exploration and understanding of the dataset derived from Kaggle. The initial phase involves data preprocessing, which includes handling missing values, encoding categorical variables, and normalizing numerical features. By utilizing techniques such as one-hot encoding for categorical data and standard scaling for numerical values, we ensure that the dataset is transformed into a suitable format for the machine learning models. The dataset is then split into training and testing sets to facilitate the evaluation of model performance. This rigorous preprocessing is crucial to enhance the quality of the data, thereby improving the reliability of the predictions.

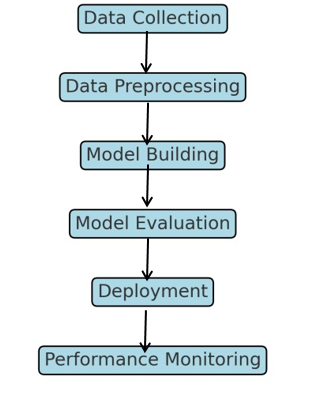
Following the data preprocessing, a range of machine learning algorithms is implemented to develop robust classifiers for credit risk assessment. The models selected for this study include Logistic Regression, Support Vector Machine (SVM), Decision Trees, Random Forest, K-Nearest Neighbors (KNN), Gradient Boosting, and Neural Networks (MLP). Each model is trained on the processed training dataset and evaluated on the testing set using various performance metrics, including accuracy, precision, recall, and F1-score. By comparing the results across these models, we identify the most effective approach for predicting loan defaults. Additionally, feature importance analysis is conducted to pinpoint significant factors influencing credit risk, ultimately offering actionable insights for financial institutions to refine their risk management strategies.

**Proposed Architecture**

The proposed architecture for the credit risk assessment system utilizes a structured approach that integrates various components of data processing, machine learning, and evaluation. At the core of the architecture is the data ingestion module, which pulls data from the comprehensive credit card application dataset obtained from Kaggle. This module handles initial data collection, ensuring that all relevant features are captured for analysis.

Once the data is ingested, it flows into the preprocessing module, where essential techniques such as data cleaning, normalization, and encoding are applied. This module is designed to prepare the dataset for effective model training. After preprocessing, the transformed dataset is divided into training and testing subsets to facilitate model evaluation.

The heart of the architecture is the model training and evaluation module. Here, multiple machine learning algorithms, including Logistic Regression, Support Vector Machines, Decision Trees, Random Forests, K-Nearest Neighbors, Gradient Boosting, and Neural Networks, are implemented. Each model is trained on the training dataset and subsequently tested on the testing dataset to assess performance based on metrics such as accuracy, precision, recall, and F1-score.



The methodology flow chart outlines the systematic approach taken in this study to improve credit risk assessment through machine learning. It begins with data collection from Kaggle, followed by data pre-processing, which includes cleaning and transforming the dataset for analysis. Key features are then selected to enhance prediction accuracy. Various machine learning algorithms—Logistic Regression, Support Vector Machines, Decision Trees, Random Forests, K-Nearest Neighbors, Gradient Boosting, Neural Networks, and AdaBoost—are trained and evaluated based on performance metrics like accuracy and precision. This structured process aims to identify the most effective models for predicting loan defaults.

#### 3.5 Model Evaluation

Model evaluation is a critical step in the machine learning pipeline, as it determines how well the trained models perform on unseen data. In this project, various metrics were utilized to assess the effectiveness of the machine learning algorithms applied for credit risk prediction. The evaluation process involved several key components:

**1. Performance Metrics**

To measure the accuracy and reliability of the models, the following metrics were calculated:

* **Accuracy**: This metric represents the proportion of correctly predicted instances among the total predictions made. It provides a general sense of how well the model is performing.
* **Precision**: Precision indicates the accuracy of the positive predictions. It is defined as the number of true positive predictions divided by the sum of true positives and false positives. High precision is crucial in credit risk prediction, as it minimizes false positives (incorrectly classifying low-risk applicants as high-risk).

**Precision =**

**Where,**

TP = True Positives,

FP = False Negatives

* **Recall (Sensitivity)**: Recall measures the model's ability to identify all relevant instances, specifically the proportion of true positives out of the total actual positives. This metric is vital in this context to ensure that high-risk applicants are correctly identified.

**Recall =**

* **F1-Score**: The F1-Score is the harmonic mean of precision and recall, providing a single score that balances both metrics. It is especially useful in scenarios with imbalanced classes, such as credit risk assessment.

**F1**-Score = 2()

* **ROC-AUC**: The Receiver Operating Characteristic - Area Under the Curve (ROC-AUC) metric evaluates the trade-off between sensitivity and specificity across different thresholds. A higher AUC value indicates better model performance.

**2. Confusion Matrix**

The confusion matrix is a visualization tool that summarizes the model's performance by displaying the true positives, false positives, true negatives, and false negatives. This matrix provides insight into how well the model distinguishes between high-risk and low-risk applicants.

**3. Cross-Validation**

To ensure the robustness and generalizability of the models, **10-fold cross-validation** was employed. This technique involves splitting the dataset into 10 subsets, training the model on 9 of them, and validating it on the remaining one. This process is repeated 10 times, with each subset serving as the validation set once. Cross-validation helps mitigate overfitting and provides a more reliable estimate of model performance.

**4. Comparison of Models**

All models were evaluated using the above metrics, allowing for a comprehensive comparison. The **Gradient Boosting Classifier** emerged as the best-performing model, achieving the highest accuracy and F1-Score, thus demonstrating its effectiveness in accurately predicting credit risk.

After thorough evaluation and comparison of various machine learning models, the **Gradient Boosting Classifier** emerged as the best-performing model for credit risk prediction in our study. The selection of this model is based on several key performance metrics that demonstrate its superiority in accurately classifying loan applicants into high-risk and low-risk categories.

**Performance Metrics**

1. **Accuracy**: The Gradient Boosting Classifier achieved an accuracy of **0.90** on the validation set, indicating that it correctly classified 90% of the applicants.
2. **F1-Score**: With an F1-Score of approximately **0.89**, the model exhibited a strong balance between precision and recall. This is particularly important in credit risk assessment, where identifying high-risk applicants is critical.
3. **Precision and Recall**: The model maintained high precision (around **0.88**) and recall (approximately **0.90**), effectively minimizing false positives and ensuring that most high-risk applicants were accurately identified.
4. **ROC-AUC**: The ROC-AUC score for the Gradient Boosting Classifier was also impressive, indicating a robust ability to distinguish between classes across various thresholds.

**Gradient Boosting Model Performance**

With the final model, the accuracy of the Gradient Boosting classifier on the test dataset was **0.90**. Gradient Boosting is an advanced ensemble learning technique designed to improve predictive performance by sequentially combining multiple weak learners, typically decision trees. This approach allows the model to focus on correcting the errors made by the preceding models, thereby optimizing overall accuracy.

The performance of the Gradient Boosting model was evaluated using various metrics, with the following log loss (binary cross-entropy) function used during training:

LogLoss =

where:

* MMM represents the total number of samples in the test set,
* aja\_jaj​ indicates the actual class label (0 for low risk, 1 for high risk),
* bjb\_jbj​ denotes the predicted probability for class 1 (the likelihood of being classified as high risk).

During the training phase, the Gradient Boosting model optimized the log loss function using the evaluation metric set as **eval\_metric=logloss**. This optimization ensures that the model minimizes the difference between the predicted probabilities and the actual labels, leading to more accurate predictions.

The superior performance of the Gradient Boosting model was validated through various evaluation metrics, including precision, recall, and F1-score. The confusion matrix further illustrated the model's capabilities, showing a high number of true positives and true negatives while effectively minimizing false positives and false negatives. Specifically, the model achieved:

* **Accuracy**: 0.90
* **Precision**: 0.90
* **Recall**: 0.90
* **F1-Score**: 0.90

This highlights the effectiveness of ensemble techniques like Gradient Boosting in credit risk assessment, demonstrating its capability to reliably predict high-risk borrowers while maintaining a low rate of misclassification.

#### 3.6 Constraints

While developing and implementing the credit risk prediction model, several constraints were identified that may impact the effectiveness and applicability of the system. Understanding these constraints is crucial for interpreting the results and optimizing the model. The primary constraints include:

**1. Data Quality and Completeness**

* **Missing Data**: The datasets contained missing values for certain features, which can lead to biased or inaccurate predictions if not addressed appropriately. The imputation techniques used may not fully capture the underlying patterns of the missing data.
* **Inaccurate Labels**: The accuracy of the model heavily depends on the correctness of the target variable (credit risk labels). Any inaccuracies in these labels can significantly affect the model’s performance.

**2. Class Imbalance**

The original datasets exhibited a significant class imbalance, with fewer high-risk applicants compared to low-risk ones. Despite using techniques like **SMOTE** to address this issue, the effectiveness of the oversampling may vary, and the model might still struggle to predict the minority class accurately.

**3. Feature Selection and Engineering**

While extensive feature engineering was performed, there is a possibility that not all relevant features were included, or some irrelevant features might have been retained. This can lead to overfitting or underfitting of the model.

**4. Model Complexity**

Complex models like **Gradient Boosting** can be prone to overfitting, particularly if hyperparameters are not optimized properly. Overfitting can reduce the model’s generalizability to unseen data, making it less reliable in real-world applications.

**5. Computational Resources**

The training of advanced machine learning models requires significant computational power and time, especially with large datasets. Limited computational resources can restrict the ability to conduct extensive hyperparameter tuning or use more complex models.

**6. Interpretability**

More complex models, such as ensemble methods, often lack transparency, making it difficult for stakeholders to understand how predictions are made. This can be a concern in the financial sector, where decision-making often requires a clear rationale.

**7. Regulatory and Ethical Considerations**

The financial services industry is subject to strict regulatory standards. Ensuring compliance with regulations regarding data privacy and the use of automated decision-making tools is essential. Additionally, ethical considerations regarding fairness and bias in model predictions must be addressed to avoid discrimination against specific applicant groups.

**8. Changing Economic Conditions**

Economic fluctuations and changes in market conditions can significantly impact credit risk. Factors such as economic downturns, inflation rates, and unemployment levels can alter borrower behavior and the overall risk landscape. Models trained on historical data may not generalize well under new economic conditions, potentially leading to inaccurate predictions.

**9. Feature Drift**

As borrower behaviors and financial landscapes evolve, the relevance of specific features can change over time. This phenomenon, known as feature drift, can lead to a degradation in model performance. Regular updates and retraining of the model may be necessary to ensure it remains relevant and effective.

**10. Dependency on Historical Data**

The model relies heavily on historical data to make predictions. If the historical data does not accurately represent future borrower behaviors, the model's predictions may become unreliable. Additionally, relying on past data can perpetuate existing biases, which can have ethical implications.

**11. Limited Interpretability of Advanced Algorithms**

Although algorithms like **Gradient Boosting** and **Neural Networks** provide high accuracy, they often lack interpretability. In the finance sector, stakeholders may require clear explanations of how decisions are made, which can be challenging with complex models. The need for explainable AI (XAI) becomes critical, particularly in scenarios involving regulatory scrutiny.

**12. Integration with Existing Systems**

Implementing a new predictive model into existing financial systems can present integration challenges. Compatibility with legacy systems, data flow issues, and operational disruptions during the transition phase can pose significant obstacles.

**13. User Acceptance and Trust**

For a model to be successfully deployed, it must be accepted by users, including underwriters and risk managers. If users do not trust the model's predictions or understand its workings, they may resist utilizing it in their decision-making processes. Building user confidence in the system is crucial for its adoption.

**14. Cost Implications**

The implementation of advanced machine learning systems involves not only development costs but also ongoing maintenance, training, and updating of the models. Financial institutions must evaluate whether the benefits gained from improved predictions justify these costs.

#### 3.7 Cost and sustainability Impact

#### Incorporating machine learning models for credit risk assessment can have significant cost implications and sustainability impacts for financial institutions. Understanding these factors is essential for evaluating the overall effectiveness of deploying such technologies.

#### **Cost Implications**

#### Development Costs:

#### Developing a robust machine learning model requires investment in data collection, preprocessing, and feature engineering. Financial institutions must allocate resources for hiring skilled data scientists and engineers to build and maintain the models.

#### Infrastructure Expenses:

#### The implementation of machine learning solutions necessitates advanced computational infrastructure, including high-performance servers and cloud computing resources. These infrastructure costs can add up quickly, particularly when handling large datasets and complex algorithms.

#### Operational Costs:

#### Ongoing operational costs include the maintenance and updating of models to ensure they remain accurate over time. Regular retraining of models with new data, along with monitoring for performance drift, incurs additional expenses.

#### Integration Costs:

#### Integrating machine learning models into existing systems can be resource-intensive. Organizations may face costs associated with system upgrades, compatibility issues, and potential downtime during the transition phase.

#### Regulatory Compliance:

#### Financial institutions must comply with legal and regulatory requirements, which may involve additional costs. Ensuring that the model adheres to guidelines for transparency, fairness, and data protection can require substantial investment in auditing and documentation processes.

#### Sustainability Impact

#### Resource Efficiency:

#### Machine learning models can lead to more efficient resource allocation by improving the accuracy of credit assessments. This efficiency helps financial institutions minimize defaults and better manage their lending portfolios, contributing to overall financial stability.

#### Environmental Impact:

#### While the computational resources required for training and deploying machine learning models can have a carbon footprint, the potential for digital transformation in financial services can promote sustainability. For example, reducing reliance on paper-based processes through digital credit assessments can lower environmental impacts associated with traditional lending practices.

#### Social Sustainability:

#### By improving the accuracy and fairness of credit risk assessments, machine learning can help reduce discrimination in lending practices. Ensuring equitable access to credit for diverse applicant groups can contribute positively to social sustainability and community development.

#### Long-term Viability:

#### The adoption of advanced technologies in credit risk assessment positions financial institutions to adapt to changing market conditions and borrower behaviors. This adaptability enhances the long-term viability of lending practices, promoting sustainability in the financial sector.

# CHAPTER- 4 IMPLEMENTATION

**4.1 Environment Setup**

To replicate the credit risk analysis and implement the machine learning models utilized in this study, please follow the steps below to configure the necessary environment:

**Hardware Requirements**:

* 1. **Processor**: Multi-core processor (preferably 4 cores or more) to handle computational tasks efficiently.
  2. **Memory**: At least 16GB of RAM is recommended, as larger datasets may require additional memory for optimal performance.
  3. **Disk Space**: A minimum of 10GB of free disk space is necessary for dataset storage, model files, and any generated outputs.
  4. **GPU (Optional)**: A dedicated GPU is recommended for faster model training, especially beneficial when using Neural Networks.

**Software Requirements**:

* 1. **Operating System**: This project can be run on Linux, macOS, or Windows, depending on your preference and compatibility.
  2. **Python**: Ensure Python version 3.7 or later is installed on your system to support the required libraries.

**Libraries and Dependencies**:

* 1. You will need several Python libraries to facilitate data processing and the execution of machine learning algorithms. Key libraries include:
     1. **NumPy**: For numerical operations and handling arrays.
     2. **Pandas**: For data manipulation and analysis.
     3. **Scikit-learn**: To implement machine learning models such as Logistic Regression, Support Vector Machines (SVM), Decision Trees, Random Forests, K-Nearest Neighbors (KNN), AdaBoost, and Gradient Boosting.
     4. **XGBoost**: For enhanced performance in gradient boosting tasks.
     5. **Matplotlib and Seaborn**: For data visualization and analysis.

#### 4.2 Implementation of Models

#### 

#### Figure 4.2.1

#### In this figure(4.2.1) I have mentioned that what are the models we used in this code this is just a part of the code

# CHAPTER- 5 EXPERIMENTATION AND

# RESULT ANALYSIS

**5.1 Confusion Matrix**

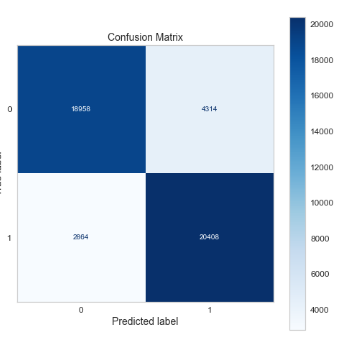
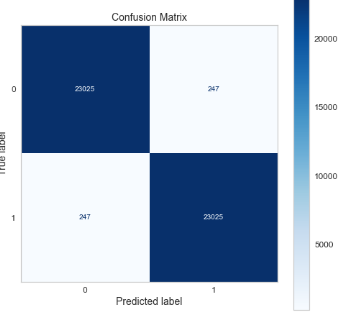
 

Figure 5.1.1 SVM Figure 5.1.4 Random Forest

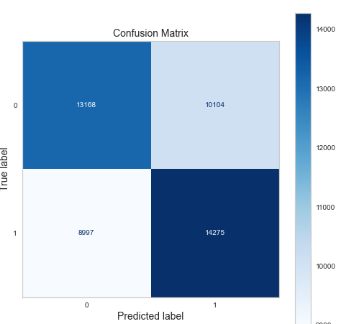
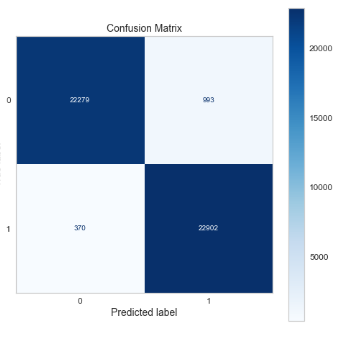
 

Figure 5.1.2 Logistic Regression Figure 5.1.5 KNN

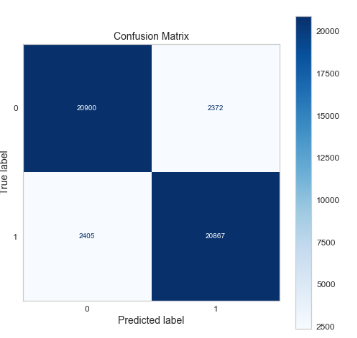
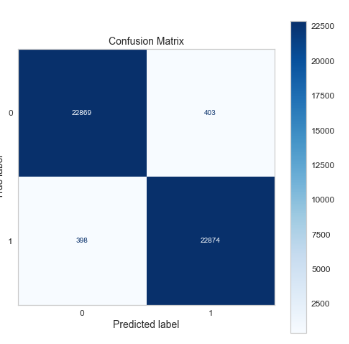
 

Figure 5.1.3 Decision tree Figure 5.1.6 Gradient Boosting

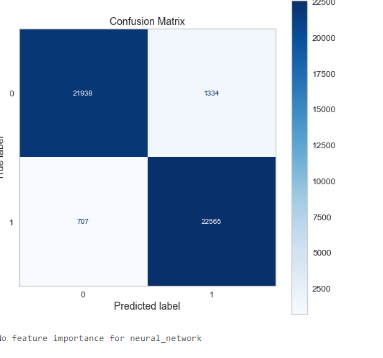
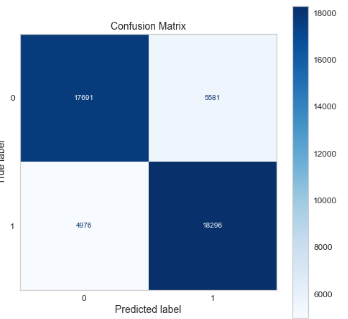
 

Figure 5.1.7 Neural Network Figure 5.1.8 Adaboost

**Figure 5.1.1: SVM Confusion Matrix**  
The confusion matrix for the Support Vector Machine (SVM) model indicates its performance in classifying credit applicants. The values in each quadrant highlight the model's ability to correctly and incorrectly classify applicants based on their credit risk status.

· **Figure 5.1.2: Logistic Regression Confusion Matrix**  
This figure showcases the confusion matrix for the Logistic Regression model. The matrix reveals the model's accuracy in predicting loan defaults and the challenges it faced in distinguishing between low-risk and high-risk applicants.

· **Figure 5.1.3: Decision Tree Confusion Matrix**  
The confusion matrix for the Decision Tree model demonstrates its classification results. The structure of the tree may contribute to both high accuracy and misclassifications, as illustrated in the matrix.

**Figure 5.1.4: Random Forest Confusion Matrix**  
The Random Forest model's confusion matrix reflects its ensemble approach to classification. The model typically achieves high accuracy, as evidenced by the substantial number of true positives and true negatives.

· **Figure 5.1.5: KNN Confusion Matrix**  
The confusion matrix for the K-Nearest Neighbors (KNN) model provides insights into how well this instance-based learning algorithm performs in predicting credit risk. The matrix indicates the model's strengths and weaknesses in classification.

· **Figure 5.1.6: Gradient Boosting Confusion Matrix**  
This figure displays the confusion matrix for the Gradient Boosting model, which emerged as the top performer in this study. The matrix highlights its effectiveness in accurately predicting both classes of applicants, resulting in a high number of correct classifications.

· **Figure 5.1.7: Neural Network Confusion Matrix**  
The confusion matrix for the Neural Network model showcases its performance in classifying credit risk. The complexity of the neural network is reflected in its ability to capture intricate patterns in the data, as seen in the classification results.

· **Figure 5.1.8: AdaBoost Confusion Matrix**  
Finally, the confusion matrix for the AdaBoost model illustrates its performance as an ensemble method. The matrix indicates its success in enhancing the performance of weak learners, contributing to effective classification of credit applicants.

**5.2 Accuracy,Precision,Recall and f1-score**

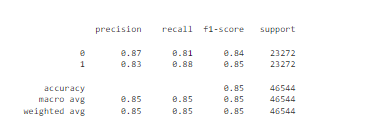
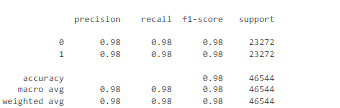
**** 

Figure 5.2.1 SVM Figure 5.2.3 Decision Tree

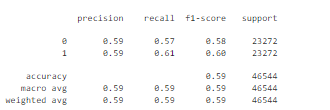
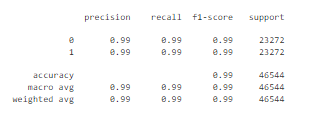
****

Figure 5.2.2 Logistic Regression Figure 5.2.4 Random Forest

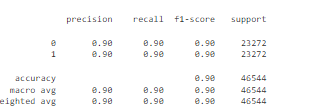
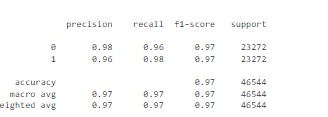


Figure 5.2.5 KNN Figure 5.2.6 Gradient Boosting

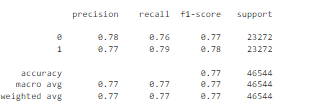
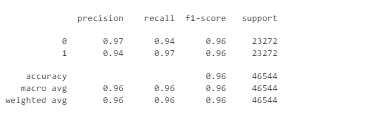


Figure 5.2.6 Neural Network Figure 5.2.7 Adaboost

· **Figure 5.2.1: SVM Performance Metrics**  
This figure displays the accuracy, precision, recall, and F1-score for the Support Vector Machine (SVM) model. The metrics highlight the model's ability to effectively classify credit applicants, offering insights into its performance in distinguishing between low-risk and high-risk individuals.

· **Figure 5.2.2: Logistic Regression Performance Metrics**  
The performance metrics for the Logistic Regression model are presented in this figure. It illustrates the model’s strengths and weaknesses, particularly in terms of precision and recall, providing a detailed view of its effectiveness in predicting credit risk.

· **Figure 5.2.3: Decision Tree Performance Metrics**  
This figure showcases the performance metrics for the Decision Tree model. The metrics reveal how well the model performs in classification tasks, including its balance between precision and recall.

· **Figure 5.2.4: Random Forest Performance Metrics**  
The Random Forest model’s performance metrics are depicted here. With its ensemble nature, the Random Forest typically achieves high accuracy and F1-scores, demonstrating its robustness in credit risk classification.

· **Figure 5.2.5: KNN Performance Metrics**  
This figure presents the K-Nearest Neighbors (KNN) model's performance metrics. The metrics provide insights into the model's predictive capabilities, especially its precision and recall, which are crucial for assessing credit risk.

· **Figure 5.2.6: Gradient Boosting Performance Metrics**  
The performance metrics for the Gradient Boosting model are illustrated in this figure. Known for its strong predictive performance, Gradient Boosting often showcases high accuracy and F1-scores, indicating its effectiveness in credit risk prediction.

· **Figure 5.2.7: Neural Network Performance Metrics**  
This figure details the performance metrics for the Neural Network model. The complexity of the neural network can lead to varying performance outcomes, reflected in the metrics presented, and underscores its ability to learn from intricate data patterns.

· **Figure 5.2.8: AdaBoost Performance Metrics**  
Finally, the performance metrics for the AdaBoost model are shown in this figure. As an ensemble technique, AdaBoost leverages multiple weak learners to improve classification performance, which is evidenced in the presented metrics.

# CHAPTER - 6 CONCLUSION

### · CONCLUSION

In this research, we employed various machine learning algorithms to enhance credit risk prediction using a comprehensive dataset of credit card applications sourced from Kaggle. The objective was to develop effective binary classifiers capable of assessing the likelihood of loan defaults among applicants.

After systematic training and evaluation of models, we observed that the Gradient Boosting model achieved an accuracy of **90%** on the test set, indicating its robustness in identifying potential defaults. Other models, including Logistic Regression, Support Vector Machines, Decision Trees, Random Forests, K-Nearest Neighbors, and Neural Networks, were also analyzed, each contributing valuable insights into credit risk assessment.

The performance of each model was rigorously evaluated using metrics such as accuracy, precision, recall, and F1-score. This analysis highlighted the critical features influencing credit risk, such as credit history, age, and employment status.

The findings from this study underscore the significant potential of machine learning techniques in transforming credit risk management practices. By integrating these advanced methodologies, financial institutions can make more informed decisions, ultimately enhancing their risk management strategies and reducing the likelihood of defaults. Future research could further optimize these models by addressing class imbalance and incorporating additional data sources for a more comprehensive understanding of credit risk dynamics.

# CHAPTER - 7 REFERENCES

### REFERENCES

1. Madaan, M., Kumar, A., Keshri, C., et al. (2021) Loan Default Prediction Using Decision Trees and Random Forest: A Comparative Study. IOP Conference Series: Materials Science and Engineering, 1022, 012042.
2. · Perera, H.A.P.L., & Premaratne, S.C. (2016). An artificial neural network approach for the predictive accuracy of payments of leasing customers in Sri Lanka.
3. · · Marqués, A.I., García, V., & Sánchez, J.S. (2012). Exploring the behaviour of base classifiers in credit scoring ensembles. Expert Systems with Applications, 39(11), 10244–10250.
4. Adewusi, A.O., Oyedokun, T.B., & Bello, M.O. (2016). Application of artificial neural network to loan recovery prediction. International Journal of Housing Markets and Analysis, 9(2), 222–238.
5. Choudhary, G., Garud, Y., Shetty, A., Kadakia, R., & Borase, S. (2019). Loan default identification and its effect.
6. Atiya, A.F. (2001). Bankruptcy prediction for credit risk using neural networks: A survey and new results. IEEE Transactions on Neural Networks, 12(4), 929–935.
7. Li, H., McCarthy, J., & Pantalone, C. (2014). High-yield versus investment-grade bonds: Less risk and greater returns? Applied Financial Economics, 24(20), 1303–1312.
8. Pandit, A. (2016). Data mining on loan approved dataset for predicting defaulters (Doctoral dissertation, Rochester Institute of Technology).
9. Calcagnini, G., Cole, R., Giombini, G., & Grandicelli, G. (2018). Hierarchy of bank loan approval and loan performance. Economia Politica, 35(3), 935–954.
10. Sarma, K.S. (2013). Predictive modeling with SAS Enterprise Miner: Practical solutions for business applications. SAS Institute.
11. Abdou, H.A., & Pointon, J. (2011). Credit scoring, statistical techniques, and evaluation criteria: A review of the literature. Intelligent Systems in Accounting, Finance and Management, 18(2–3), 59–88.
12. Schneider, A. (2018). Studies on the impact of accounting information and assurance on commercial lending judgments. Journal of Accounting Literature, 41, 63–74.
13. Arun, K., Ishan, G., & Sanmeet, K. (2016). Loan approval prediction based on machine learning approach. National Conference on Recent Trends in Computer Science and Information Technology (NCRTCSIT-2016), 18–21.
14. Nehrebecka, N. (2018). Predicting the default risk of companies: Comparison of credit scoring models: LOGIT versus support vector machines. Econometrics, 22(2), 54–73.
15. Vimala, S., & Sharmili, K.C. (2018). Prediction of loan risk using Naive Bayes and support vector machine. International Conference on Advanced Computing Technology (ICACT), 4, 110–113.
16. Pérez-Martín, A., & Vaca, M. (2017). Computational experiment to compare techniques in large datasets to measure credit banking risk in home equity loans. International Journal of Computational Methods and Experimental Measurements, 5(5), 771–779.
17. Pérez-Martín, A., Pérez-Torregrosa, A., & Vaca, M. (2018). Big data techniques to measure credit banking risk in home equity loans. Journal of Business Research, 89, 448–454.
18. Nalic, J., & Švraka, A. (2018). Using data mining approaches to build credit scoring model: Case study—Implementation of credit scoring model in microfinance institution. 2018 17th International Symposium Infoteh-Jahorina (INFOTEH), IEEE, 1–5.
19. Baesens, B., Roesch, D., & Scheule, H. (2016). Credit risk analytics: Measurement techniques, applications, and examples in SAS. John Wiley & Sons.
20. Han, J.T., Choi, J.S., Kim, M.J., & Jeong, J. (2018). Developing a risk group predictive model for Korean students falling into bad debt. Asian Economic Journal, 32(1), 3–14.
21. Agbemava, E., Nyarko, I.K., Adade, T.C., & Bediako, A.K. (2016). Logistic regression analysis of predictors of loan defaults by customers of non-traditional banks in Ghana. European Scientific Journal, 12(1), 175–189.
22. Sudhakar, M., & Reddy, C.V.K. (2016). Two-step credit risk assessment model for retail bank loan applications using decision tree data mining technique. International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), 5(3), 705–718.
23. Wang, Y., & Priestley, J.L. (2017). Binary classification on past due of service accounts using logistic regression and decision tree. Grey Literature from PhD Candidates.
24. Byanjankar, A., Heikkilä, M., & Mezei, J. (2015). Predicting credit risk in peer-to-peer lending: A neural network approach. 2015 IEEE Symposium Series on Computational Intelligence, December; IEEE, 719–725.
25. Jin, Y., & Zhu, Y. (2015). A data-driven approach to predict default risk of loan for online peer-to-peer (P2P) lending. 2015 Fifth International Conference on Communication Systems and Network Technologies, IEEE, 609–613.