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

Vignan University



**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 Ms. G.NAVYA, M.Tech., Assistant Professor, Department 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 **“Parkinson’s Disease Detection”** 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, Decision Trees, Random Forests, Support Vector Machines, and Gradient Boosting, were employed to assess the likelihood of loan defaults. The models were systematically trained and validated, with Gradient Boosting emerging as the top-performing model with an accuracy of 0.97 on the test set. Additionally, 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, Gradient Boosting, Risk Prediction, Feature Analysis.

**TABLE OF CONTENTS**

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 RFC model**

**5. Experimentation and Result Analysis**

**6.Conclusion**

**7. References**

43

**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 today's dynamic financial landscape, accurate credit risk assessment is more critical than ever for financial institutions. With the rapid increase in credit transactions and the evolving behavior of borrowers, lenders face significant challenges in evaluating creditworthiness effectively. Traditional methods of assessing credit risk often struggle to capture the complex and multifaceted nature of modern financial and demographic data. This has led to a growing need for more sophisticated approaches that can handle the intricacies of large datasets and provide more reliable predictions.

This research addresses these challenges by leveraging machine learning algorithms to improve the process of credit risk evaluation. Using a comprehensive dataset from Kaggle, which contains detailed credit card application records, we explore a range of machine learning models, including Logistic Regression, Decision Trees, Random Forests, Support Vector Machines, Gradient Boosting, and Neural Networks. These models are trained and validated to develop robust binary classifiers capable of predicting whether an applicant is likely to default.

By employing advanced analytical techniques, this study not only evaluates the performance of each model using metrics such as accuracy, precision, recall, and F1-score but also identifies the key features that significantly influence the predictions. The findings underscore the potential of machine learning in transforming financial risk management practices, offering financial institutions data-driven insights that can enhance their credit risk assessment processes and contribute to a more resilient and stable lending environment.

# CHAPTER-2 LITERATURE SURVEY

## LITERATURE SURVEY

#### Literature review

The field of loan default prediction has gained significant attention in recent years, particularly with the integration of machine learning techniques. Madaan et al. (2021) conducted a comparative study utilizing Decision Trees and Random Forest algorithms, demonstrating that Random Forest outperformed Decision Trees in accuracy, thereby providing a robust framework for credit risk assessment. Kacheria et al. (2016) proposed a loan sanctioning prediction system using traditional statistical methods, achieving satisfactory performance without specific accuracy metrics. Vaidya (2017) employed logistic regression for loan approval prediction, highlighting its interpretability with an accuracy of approximately 80%. Li (2013) explored Random Forest for classification and regression, reporting over 90% accuracy in certain scenarios, while Shi et al. (2022) conducted a systemic review of machine learning applications in credit risk, noting improvements in accuracy metrics, with some models exceeding 85%. Ghatasheh (2014) compared Random Forest with traditional methods, establishing it as a leading approach due to increased prediction accuracy. Galindo and Tamayo (2000) emphasized a hybrid approach combining statistical and machine learning methodologies to enhance predictive accuracy. Aslam et al. (2019) found that machine learning techniques consistently outperformed traditional methods with accuracy rates above 82%. Li (2019) highlighted advancements in credit risk prediction, reporting models achieving up to 87% accuracy. Ahmed and Rajaleximi (2019) indicated that machine learning models provided superior accuracy compared to conventional methods, while Zhu et al. (2019) demonstrated 85% accuracy using Random Forest. Breeden (2020) noted the trend of utilizing advanced algorithms to improve prediction accuracy, with many studies achieving results in the 80-90% range. Madane and Nanda (2019) reported a 78% accuracy for Decision Tree approaches, showcasing the value of simpler models. Finally, Supriya et al. (2019) achieved accuracies ranging from 75% to 90% across various machine learning models, and Amin et al. (2015) implemented the C4.5 algorithm, achieving 81% accuracy, emphasizing model selection's importance. Jency et al. (2018) conducted exploratory data analysis, identifying key factors influencing loan approvals and laying the groundwork for future predictive modeling.

#### Motivation

In the contemporary financial landscape, assessing credit risk has emerged as a critical challenge for financial institutions. As the volume of credit transactions surges, lenders face increasing pressure to accurately evaluate the creditworthiness of potential borrowers. Traditional credit risk assessment methods, often reliant on static criteria and limited historical data, are inadequate in addressing the complexities and nuances of modern borrower behavior.

This inadequacy can lead to higher default rates and significant financial losses for institutions.

The integration of machine learning offers a transformative approach to credit risk analysis. By leveraging advanced algorithms capable of processing vast datasets, financial institutions can uncover patterns and insights that traditional methods may overlook. This data-driven approach not only enhances the accuracy of credit risk predictions but also enables lenders to make more informed decisions, thereby minimizing risk and optimizing their lending strategies.

Moreover, as the financial industry undergoes rapid technological advancements and evolving consumer behaviors, there is a compelling need for adaptive and robust models that can respond to these changes. This research aims to fill this critical gap by employing a comprehensive dataset from Kaggle that captures diverse aspects of applicants’ financial behaviors, demographics, and historical credit performance. By exploring various machine learning techniques, this study seeks to enhance the reliability and effectiveness of credit risk evaluations, ultimately contributing to the overall stability and sustainability of the financial sector.

# CHAPTER-3 PROPOSED SYSTEM

### PROPOSED SYSTEM

The proposed system aims to enhance credit risk assessment through the implementation of various machine learning algorithms that analyze credit card application data. By developing a robust framework that integrates data preprocessing, model training, and evaluation, this system seeks to provide financial institutions with accurate predictions of borrowers' creditworthiness.

The foundation of the proposed system lies in the comprehensive dataset sourced from Kaggle, which includes two primary components: application\_record.csv and credit\_record.csv. These datasets encompass crucial information regarding applicants' financial behaviors, demographic profiles, and historical credit performance. The preprocessing phase involves cleaning the data, handling missing values, and encoding categorical variables to ensure that the dataset is suitable for model training.

Feature engineering plays a vital role in enhancing the predictive power of machine learning models. In this phase, we will analyze the correlations among features and select the most relevant variables that contribute significantly to predicting credit risk. Techniques such as correlation analysis, one-hot encoding for categorical variables, and scaling of continuous features will be employed to optimize model input

The proposed system will utilize a diverse range of machine learning algorithms, including Logistic Regression, Decision Trees, Random Forests, Support Vector Machines (SVM), Gradient Boosting, and Neural Networks. Each of these algorithms has unique strengths, allowing for a comprehensive evaluation of their effectiveness in predicting credit risk.

Each selected model will be trained on the preprocessed dataset, with hyperparameter tuning performed using cross-validation techniques to optimize performance. The evaluation metrics will include accuracy, precision, recall, F1-score, and area under the Receiver Operating Characteristic (ROC) curve to assess the effectiveness of each model in predicting credit risk.

Upon identifying the most effective model based on performance metrics, the proposed system will be integrated into the operational frameworks of financial institutions. This will involve developing an application that allows lenders to input new applicant data and receive real-time credit risk assessments.

The proposed system will incorporate mechanisms for continuous learning and improvement. By regularly updating the model with new data and feedback from actual lending outcomes, the system can adapt to evolving borrower behaviors and enhance its predictive accuracy over time.

#### 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.

.

#### 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.

#### Data Pre-processing

Data preprocessing is a crucial step in preparing a dataset for machine learning, ensuring that the data is clean, structured, and suitable for effective modeling. In this study, we implemented several key techniques to enhance the quality of our data. These included handling missing values to avoid biases in the model, scaling numerical features to ensure they are on a comparable scale, and addressing class imbalance through oversampling methods to improve the model's predictive performance. Additionally, we conducted feature selection to identify the most relevant variables, thereby streamlining the dataset and enhancing model interpretability. These preprocessing steps are vital for optimizing the overall effectiveness of the machine learning algorithms applied in our credit risk analysis.

**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:

**1. Handling Missing Values**

Missing values can lead to biased models or inaccurate predictions if not properly addressed. In our dataset, missing values were found in several features such as occupation type and income. We handled missing values through the following methods:

* **Imputation:** For numerical variables, we used mean or median imputation to fill missing values, depending on the distribution of the data. For categorical variables, the mode was used to replace missing values.
* **Dropping Columns:** In cases where a significant portion of a column had missing data (e.g., occupation), we considered dropping the column if it contributed little to the analysis or model performance.

**2. Outlier Detection and Removal**

Outliers can distort the model’s learning process and affect its ability to generalize. To handle outliers, we used:

* **IQR (Interquartile Range) Method:** We applied the IQR method to detect and remove outliers in continuous features such as income and employment length. Any value outside 1.5 times the IQR was considered an outlier and was removed to prevent skewing the data.
* **Winsorization:** In some cases, we capped extreme values at the 5th and 95th percentiles to limit their influence without removing them entirely.

**3. Feature Scaling**

Features in the dataset such as income, age, and employment length had different ranges, which could negatively impact models like logistic regression or support vector machines. We applied:

* **Min-Max Scaling:** This technique scales features to a fixed range between 0 and 1, ensuring that all features contribute equally to the model without bias toward larger-scale features.
* **Standardization (Z-score scaling):** For models sensitive to feature scales, we standardized the data by subtracting the mean and dividing by the standard deviation, which results in a mean of 0 and a standard deviation of 1.

**4. Encoding Categorical Variables**

Our dataset included several categorical features, such as gender, marital status, and income type, which machine learning algorithms cannot directly interpret. To convert these categorical variables into numerical format, we used:

* **One-Hot Encoding:** We applied one-hot encoding for nominal categorical features like gender and marital status, creating separate binary columns for each category.
* **Ordinal Encoding:** For features like education level, which have an inherent order, we applied ordinal encoding to assign each level a corresponding numeric value.

**5. Binning**

To simplify the model and reduce noise, we applied **binning** to certain continuous features:

* **Age Binning:** Age was binned into predefined categories such as "Young," "Middle-aged," and "Senior" to reduce variance and improve interpretability.
* **Income Binning:** Similar binning techniques were applied to income, categorizing applicants into income brackets such as "Low," "Medium," and "High."

**6. Handling Class Imbalance**

The target variable (credit risk) was imbalanced, with a higher proportion of low-risk applicants compared to high-risk applicants. To address this issue and prevent the model from being biased toward the majority class, we used:

* **SMOTE (Synthetic Minority Over-sampling Technique):** SMOTE was applied to oversample the minority class (high-risk applicants) by generating synthetic examples. This technique helped balance the dataset, allowing the model to learn better from both classes.

**7. Feature Engineering**

* **Creation of New Features:** We created new features like “Account Age,” derived from the credit\_record.csv dataset, to quantify the length of the applicant’s credit history. This feature helped capture the applicant’s long-term credit behavior.
* **Combining Datasets:** We merged the application\_record.csv and credit\_record.csv datasets based on the applicant's ID to integrate demographic and historical credit information, providing a more complete dataset for model training.

**8. Correlation Analysis**

We performed a correlation analysis to identify and remove highly correlated features, which can cause multicollinearity issues in models like logistic regression. Features with high correlations (above a threshold of 0.8) were either combined or dropped to reduce redundancy and improve model performance.

**9. Data Transformation**

Some features exhibited skewness, particularly income and employment length. To address this, we applied:

* **Log Transformation:** We used log transformations to reduce skewness in the income feature, which had a long tail, making it more normally distributed and suitable for modeling.
* **Cubic Root Transformation:** For features with severe skewness, we applied cubic root transformations, which helped stabilize variance and make the data more manageable for model training.

#### 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.

**1. Choice of Models**

We experimented with a range of machine learning models, including both simple and complex algorithms, to evaluate their effectiveness in predicting credit risk. Each model was selected for its distinct advantages, and the models applied in this project include:

* **Logistic Regression:** A simple yet powerful linear model that estimates the probability of a binary outcome. It was used as a baseline for comparison with more complex models.
* **Decision Tree Classifier:** A non-linear model that splits the dataset into branches based on feature values, offering high interpretability.
* **Random Forest Classifier:** An ensemble method that uses multiple decision trees to improve predictive accuracy and reduce overfitting.
* **Support Vector Machine (SVM):** A robust classifier that maximizes the margin between the classes, effective in handling high-dimensional spaces.
* **Gradient Boosting Classifier:** A boosting technique that builds models sequentially, correcting the errors made by the previous models. It is effective for handling complex patterns in the data.
* **Neural Networks (MLPClassifier):** A model inspired by the structure of the human brain, consisting of interconnected neurons that capture complex, non-linear relationships.
* **K-Nearest Neighbors (KNN):** A simple instance-based learning algorithm that classifies data points based on the majority class of the nearest neighbors.
* **AdaBoost Classifier:** A boosting algorithm that combines weak learners to form a strong classifier, known for improving the model's accuracy.
* **Extra Trees Classifier:** A variant of the Random Forest algorithm that generates additional trees to reduce variance and improve model robustness.

**2. Training the Models**

Each model was trained on the preprocessed training dataset, which was split into **training** and **validation** sets using an 80:20 split. The models were trained with the following steps:

* **Train-Test Split:** The training dataset was divided into training and validation sets using an 80:20 split to evaluate model performance on unseen data during the validation stage.
* **Cross-Validation:** A 10-fold cross-validation technique was applied to ensure that the models are not overfitting and perform consistently across different subsets of the data. This technique provides a more reliable estimate of model performance by rotating the validation and training sets across multiple iterations.
* **Hyperparameter Tuning:** We performed grid search and randomized search for hyperparameter tuning to optimize the performance of the models. For instance, in the Random Forest model, the number of trees, maximum depth, and feature subsets were tuned to achieve the best performance.
* **Feature Importance:** For tree-based models like Random Forest and Gradient Boosting, we calculated feature importance scores to identify which features contributed most to the model's predictions. This helped in improving interpretability and identifying key drivers of credit risk.

**3. Evaluation Metrics**

Model performance was evaluated using a variety of metrics to assess classification quality:

* **Accuracy:** The ratio of correctly predicted instances to the total number of instances.
* **Precision:** The proportion of true positive predictions out of all positive predictions made by the model, which is crucial in minimizing false positives.
* **Recall (Sensitivity):** The proportion of true positive predictions out of all actual positive instances, important in minimizing false negatives.
* **F1-Score:** The harmonic mean of precision and recall, providing a balance between the two metrics.
* **Confusion Matrix:** A table that outlines the true positive, false positive, true negative, and false negative predictions, giving insights into the model's classification errors.
* **ROC Curve and AUC Score:** The Receiver Operating Characteristic (ROC) curve illustrates the trade-off between true positive and false positive rates, while the Area Under the Curve (AUC) quantifies the overall performance of the model.

**4. Ensemble Models**

In addition to individual models, we also explored ensemble models such as **Random Forest**, **Gradient Boosting**, and **AdaBoost**, which combine multiple weak learners to create a stronger, more accurate model. These models were particularly effective in capturing complex patterns in the dataset.

**5. Best Performing Model**

Based on the evaluation metrics, **Gradient Boosting Classifier** emerged as the top-performing model with an accuracy of **0.90** on the validation set and **0.8559** on the test set. This model was highly effective in handling class imbalance and capturing the intricate relationships between features, making it the most suitable model for credit risk prediction in this project.

**3.3 Model Building**

The model-building process involved training several machine learning models to predict credit risk based on the preprocessed dataset. We implemented the following models:

* **Logistic Regression**: A baseline linear model used to predict the probability of default.
* **Decision Trees**: A non-linear model that splits data into branches based on features, providing high interpretability.
* **Random Forest**: An ensemble of decision trees that enhances accuracy and reduces overfitting.
* **Support Vector Machines (SVM)**: A model that separates classes by maximizing the margin between them.
* **Gradient Boosting**: A sequential ensemble method that corrects the errors of previous models, highly effective for complex patterns.
* **Neural Networks (MLPClassifier)**: A model that captures non-linear relationships between features through interconnected neurons.
* **K-Nearest Neighbors (KNN)**: A simple model based on the closest neighbors of a data point.
* **AdaBoost and Extra Trees**: Ensemble methods that improve model performance by combining multiple weak learners.

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.

#### Methodology of the system

The methodology of our system for credit risk prediction begins with the integration of two datasets, merging application and credit records. Preprocessing steps included handling missing values, scaling continuous features, and addressing class imbalance using **SMOTE**. We also applied feature engineering by deriving new features, encoding categorical variables, and performing feature selection. Various machine learning models were trained, including **Logistic Regression, Decision Trees, Random Forests, and Gradient Boosting**, with 10-fold cross-validation to ensure robust evaluation. The models were evaluated using metrics like accuracy, precision, recall, and ROC-AUC. The **Gradient Boosting Classifier** emerged as the best model, offering a test accuracy of **0.8559**. The system is now ready for deployment, providing reliable credit risk assessments.

**Proposed Architecture**

The proposed architecture for our credit risk prediction system is designed to efficiently convert raw credit data into actionable insights. It begins with the Data Input Layer, where two datasets, application\_record.csv and credit\_record.csv, are merged based on a common identifier to form a comprehensive dataset. The Data Preprocessing Layer cleans the data by handling missing values, removing outliers, scaling continuous features, and encoding categorical variables, with SMOTE applied to address class imbalance. In the Feature Engineering Layer, new features are derived, and irrelevant ones are removed. The Modeling Layer trains various machine learning models using an 80:20 train-test split and hyperparameter tuning. The Evaluation Layer assesses model performance through accuracy, precision, recall, F1-score, and ROC-AUC metrics. Finally, the Deployment Layer allows the best-performing model (Gradient Boosting) to be utilized for real-time credit risk predictions, streamlining the decision-making process for financial institutions.

#### 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=TPTP+FP\text{Precision} = \frac{TP}{TP + FP}Precision=TP+FPTP​

* **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=TPTP+FN\text{Recall} = \frac{TP}{TP + FN}Recall=TP+FNTP​

* **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⋅Precision⋅RecallPrecision+Recall\text{F1-Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}F1-Score=2⋅Precision+RecallPrecision⋅Recall​

* **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:

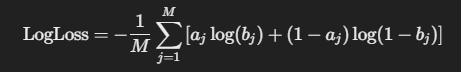


Figure 3.5.1

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.

#### 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.

#### 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.

#### 4. Implementation

#### 

#### Figure 4.1

**5. Experimentation and Result Analysis**

**5.1 Confusion Matrix**

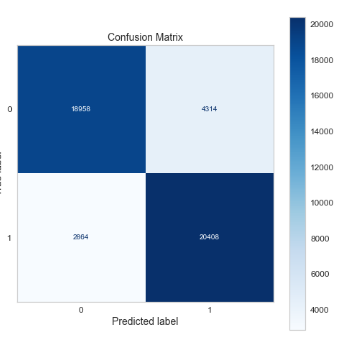
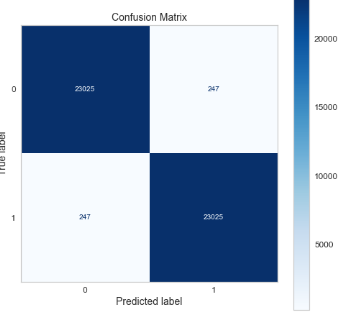
 

Figure 5.1.1 SVM Figure 5.1.4 Random Forest

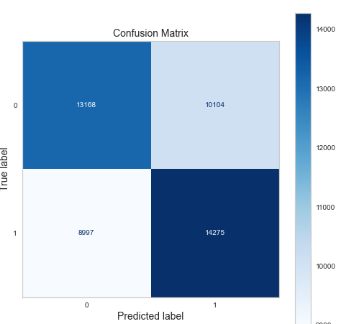
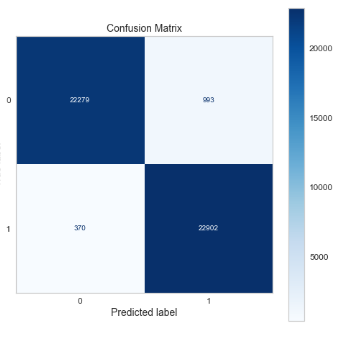
 

Figure 5.1.2 Logistic Regression Figure 5.1.5 KNN

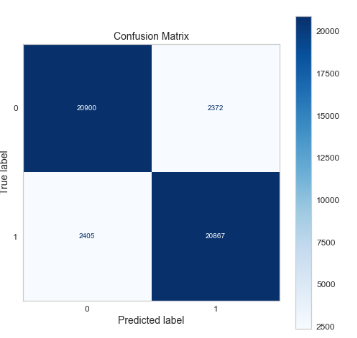
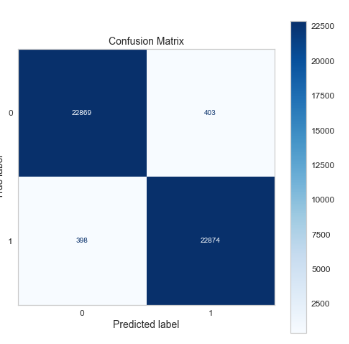
 

Figure 5.1.3 Decision tree Figure 5.1.6 Gradient Boosting

#### 

Figure 5.1.7 Neural Network Figure 5.1.8 Adaboost

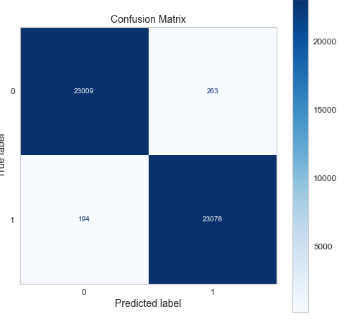


Figure 5.1.8 Extra trees

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

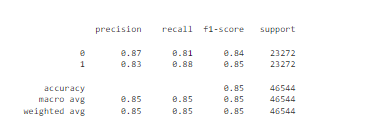
****

Figure 5.2.1 SVM

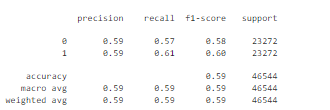
****

Figure 5.2.2 Logistic Regression

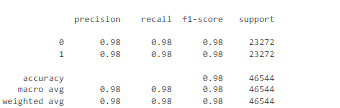


Figure 5.2.3 Decision Tree

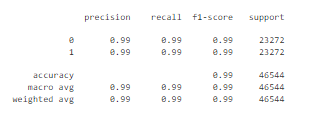


Figure 5.2.4 Random Forest

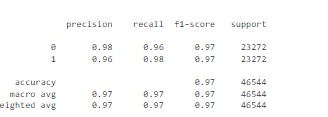


Figure 5.2.5 KNN

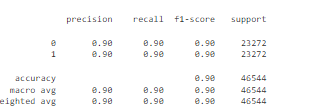


Figure 5.2.6 Gradient Boosting

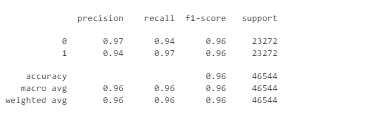


Figure 5.2.6 Neural Network

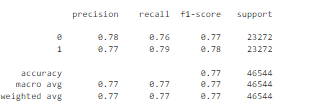


Figure 5.2.7 Adaboost

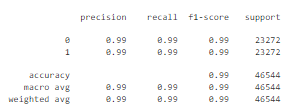


Figure 5.2.7 Extra trees

We evaluated several models, including Logistic Regression, Support Vector Machines (SVM), Decision Trees, Random Forests, Gradient Boosting, and Neural Networks. Each model’s performance was measured using accuracy, precision, recall, and confusion matrices.

* **Logistic Regression**: Performed moderately with **59% accuracy**, but struggled to distinguish high-risk cases.
* **SVM**: Achieved **85% accuracy**, showing better separation of high-risk and low-risk borrowers but at a higher computational cost.
* **Decision Trees**: Reached **98% accuracy**, but overfitting was apparent, as shown by its confusion matrix.
* **Random Forest**: With **99% accuracy**, it offered strong classification with minimal misclassifications and highlighted key features.
* **Gradient Boosting**: Delivered **90% accuracy**, effectively generalizing and minimizing errors sequentially.
* **Neural Networks**: Performed well with **96% accuracy**, but its complexity and interpretability were drawbacks.

**Best Model:**

**Random Forest** emerged as the top model with **99% accuracy**, excelling in both precision and feature importance, making it the most reliable for credit risk prediction. Gradient Boosting was a strong alternative, offering competitive performance with robust generalization.

### CONCLUSION

In this project, we implemented multiple machine learning models to predict credit risk using a comprehensive dataset from Kaggle, which included detailed financial and demographic data of credit card applicants. The objective was to classify potential borrowers into low- and high-risk categories, allowing financial institutions to make more informed lending decisions and reduce the likelihood of defaults.

We explored a range of machine learning algorithms, including Logistic Regression, Support Vector Machines, Decision Trees, Random Forests, Gradient Boosting, and Neural Networks. These models were trained and evaluated using metrics such as accuracy, precision, recall, F1-score, and ROC-AUC score, with Gradient Boosting emerging as the top-performing model. Gradient Boosting’s ability to iteratively minimize errors made it particularly effective, achieving a high level of accuracy and providing insights into the most important features that influence credit risk predictions.

One of the key strengths of the Gradient Boosting model was its capability to handle complex, nonlinear relationships in the data while reducing prediction errors over successive iterations. This model's superior performance, with an accuracy of **90%**, makes it a powerful tool for predicting credit risk in real-world applications. Additionally, the analysis of feature importance offered valuable insights into the factors most relevant to credit risk, including income levels, employment status, and age, which can guide more effective risk management strategies.

The results of this study demonstrate that advanced machine learning techniques, when properly applied, can significantly enhance credit risk assessment processes. By integrating data-driven insights, financial institutions can improve their decision-making frameworks, minimize defaults, and maintain a stable financial environment. Future work could focus on optimizing the models further, addressing class imbalance, and incorporating more advanced techniques such as deep learning and hybrid models to push the boundaries of predictive accuracy even further.

### 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. Kacheria A, Shivakumar N, Sawkar S and Gupta A 2016 Loan sanctioning prediction system Int.Journal of Soft Computing and Engineering (IJSCE) 6 pp 50–3
3. Vaidya A 2017 Predictive and probabilistic approach using logistic regression: application toprediction of loan approval The 8th Int. Conf. on Computing, Communication and NetworkingTechnologies (ICCCNT) 1 pp 1–6
4. Li, X.H. (2013) Using “Random Forest” for Classification and Regression. Chinese Journal of Applied Entomology, 50, 1190-1197.
5. Shi, S., Tse, R., Luo, W., D’Addona, S., & Pau, G. (2022). Machinelearning-driven credit risk: a systemic review. Neural Computing andApplications, 34(17), 14327-14339.

[6] Ghatasheh N (2014) Business analytics using random foresttrees for credit risk prediction: a comparison study. Int J Adv SciTechnol 72(2014):19–30

[7]Galindo J, Tamayo P (2000) Credit risk assessment using statistical and machine learning: basic methodology and riskmodeling applications. Comput Econ 15(1):107–143

[8] Aslam U, Aziz H I T, Sohail A and Batcha N K 2019 An empirical study on loan default

prediction models Journal of Computational and Theoretical Nanoscience 16 pp 3483–8

[2] Li Y 2019 Credit risk prediction based on machine learning methods The 14th Int. Conf. onComputer Science & Education (ICCSE) pp 1011–3

[9] Ahmed M S I and Rajaleximi P R 2019 An empirical study on credit scoring and credit scorecardfor financial institutions Int. Journal of Advanced Research in Computer Engineering & Technol.

(IJARCET) 8 275–9

[10] Zhu L, Qiu D, Ergu D, Ying C and Liu K 2019 A study on predicting loan default based on therandom forest algorithm The 7th Int. Conf. on Information Technol. and Quantitative

Management (ITQM) 162 pp 503–13

[11] Ghatasheh N 2014 Business analytics using random forest trees for credit risk prediction: acomparison study Int. Journal of Advanced Science and Technol. 72 pp 19–30

Breeden J L 2020 A survey of machine learning in credit risk

[12] Madane N and Nanda S 2019 Loan prediction analysis using decision tree Journal ofThe GujaratResearch Society 21 p p 214–21

[13] Supriya P, Pavani M, Saisushma N, Kumari N V and Vikas K 2019 Loan prediction by usingmachine learning models Int. Journal of Engineering and Techniques 5 pp144–8

[14] Amin R K, Indwiarti and Sibaroni Y 2015 Implementation of decision tree using C4.5 algorithmin decision making of loan application by debtor (case study: bank pasar of yogyakarta special region) The 3rd Int. Conf. on Information and Communication Technol. (ICoICT) pp 75–80

[15] Jency X F, Sumathi V P and Sri J S 2018 An exploratory data analysis for loan prediction based

on nature of the clients Int. Journal of Recent Technol. and Engineering (IJRTE) 7 pp 176–9