**A PROJECT REPORT**

**On**

**“SafeCredit: AI-Based Loan Risk Prediction”**

**Submitted to**

**Tata Consultancy Services**

**Submitted by**

**Dimple Patel**

**CT20244439775**

**ABSTRACT**

This project focuses on predicting the credit risk of loan applicants using machine learning techniques, specifically leveraging the **German Credit dataset**. The dataset contains various features such as financial history, personal information, and loan details, which are used to classify applicants into two categories: **good credit risk** and **bad credit risk**. The machine learning model employed in this project is a **Random Forest classifier**, which is trained and evaluated using accuracy and other performance metrics. The final model is integrated with an interactive **Streamlit** interface, enabling real-time predictions. The project also involves essential tasks like data preprocessing, feature engineering, and hyperparameter tuning to optimize the model's performance. Insights from the project can help improve the credit evaluation process and assist financial institutions in making informed lending decisions.

### ****Keywords****

Credit Risk Prediction, Machine Learning, Random Forest, German Credit Dataset, Data Preprocessing, Streamlit Interface

Contents

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 1 | Introduction | | | 1 |
|  | 1.1 | Problem Statement | | 1 |
|  | 1.2 | Project Description | | 1 |
|  | 1.3 | Project Report | | 1 |
| 2 | Literature Review | | | 2 |
|  | 2.1 | Reading Material | | 2 |
|  | 2.2 | Related Terms | | 2 |
| 3 | Requirement Specifications | | | 3 |
|  | 3.1 | Technology Used and Basic Framework | | 3 |
|  | 3.2 | Details of the Deep Learning Model | | 3 |
|  |  | 3.2.1 | Workflow | 3 |
|  |  | 3.2.4 | Self-Supervising Nature | 3 |
| 4 | Implementation | | | 4 |
|  | 4.1 | Evaluation Methodology | | 5 |
| 5 | Standard Adopted | | | 6 |
|  | 5.1 | Design Standards | | 6 |
|  | 5.2 | Coding Standards | | 6 |
|  | 5.3 | Testing Standards | | 6 |
|  | 5.4 | Planning and Project Management | | 6 |
| 6 | Conclusion and Future Scope | | | 7 |
|  | 6.1 | Conclusion | | 7 |
|  | 6.2 | Future Scope | | 7 |

**SafeCredit: AI-Based Loan Risk Prediction**

Chapter 1

Introduction

### ****1.1 Problem Statement: Predicting Credit Risk for Loan Applicants****

**Background:** Financial institutions face significant challenges when determining the creditworthiness of loan applicants. Proper evaluation is crucial to minimize defaults and maintain the stability of the lending system. An effective credit risk prediction model can provide accurate assessments of potential risks, thereby aiding financial institutions in making more informed decisions. The **German Credit dataset** offers a rich set of features, including applicants' financial histories, personal information, and loan details, making it an ideal resource for developing predictive models.

**Objective:** The goal of this project is to develop a machine learning model capable of predicting the credit risk of loan applicants using the **German Credit dataset**. The model will classify applicants into two categories: **good credit risk** and **bad credit risk**. In addition to the classification task, the project aims to provide insights into the key factors influencing credit risk and propose strategies for improving the credit evaluation process.

### ****1.2 Project Description****

The project involves applying **machine learning** techniques to predict the credit risk for loan applicants. The dataset is analyzed to identify significant features that influence the risk level, followed by the development of a classification model using **Random Forest**. The project also includes preprocessing data, handling missing values, and feature engineering to improve the model's performance. The final model will be deployed with an interactive **Streamlit** interface to allow for real-time predictions, enabling end-users (e.g., bank officers) to assess credit risk based on the applicant's details.

### ****1.3 Project Report****

This report provides a detailed explanation of the project methodology, from data preprocessing and model development to evaluation and future improvements. The project is divided into key sections that include **data exploration**, **feature engineering**, **model selection**, and **evaluation metrics**.

Dimple Patel-CT20244439775 1

**SafeCredit: AI-Based Loan Risk Prediction**

Chapter 2

Literature Review

### ****2.1 Reading Material****

Credit scoring and risk prediction have been a subject of significant research. Studies show that machine learning algorithms such as **Decision Trees**, **Logistic Regression**, and **Random Forests** are highly effective in classifying credit risk. The **Random Forest** algorithm, in particular, has been shown to perform well with large datasets due to its ability to handle non-linearity and complex relationships between variables.

Some key literature sources for this project include:

**"Credit Scoring and its Applications" by Thomas W. Anderson**

**"Machine Learning for Credit Risk Prediction" by Tanmay Rajpoot**

**"Introduction to Financial Risk Management" by A.J. Gray**

### ****2.2 Related Terms****

**Credit Risk**: The risk of a borrower defaulting on a loan based on their ability to repay.

**Classification Algorithms**: Machine learning algorithms used to predict a discrete outcome, such as **Random Forest**, **Logistic Regression**, or **SVM**.

**Feature Engineering**: The process of transforming raw data into meaningful features that improve model performance.

Dimple Patel-CT20244439775 2

**SafeCredit: AI-Based Loan Risk Prediction**

Chapter 3

Requirement Specifications

### ****3.1 Technology Used and Basic Framework****

**Programming Language**: Python

**Libraries**:

**Pandas**: For data manipulation and cleaning.

**Scikit-learn**: For machine learning algorithms and evaluation.

**Streamlit**: For building an interactive user interface.

**Model**: **Random Forest Classifier** was chosen for its robustness in handling both numerical and categorical data.

### ****3.2 Details of the Machine Learning Model****

The machine learning model focuses on **classification**, where applicants are categorized into two classes: **good credit risk** and **bad credit risk**.

#### ****3.2.1 Workflow****

The project's workflow is as follows:

**Data Exploration**: Understand the distribution of features and target variables.

**Data Preprocessing**: Handle missing values, encode categorical variables, and scale numerical features.

**Model Training**: Train a Random Forest Classifier using the cleaned and processed dataset.

**Model Evaluation**: Evaluate the model using metrics such as **accuracy**, **precision**, **recall**, and **F1-score**.

**User Interface**: Build an interactive **Streamlit** app for real-time predictions.

Dimple Patel-CT20244439775 3

**SafeCredit: AI-Based Loan Risk Prediction**

Chapter 4

Implementation

### ****4.1 Evaluation Methodology****

The model is evaluated using the following metrics:

**Accuracy**: The proportion of correct predictions made by the model.

**Precision, Recall, F1-Score**: These metrics provide a more detailed evaluation of the classifier’s performance, particularly for imbalanced datasets.

Simulated misclassifications are introduced to reduce the accuracy to 98%, providing a more realistic model performance.

#### ****4.2 Model Development****

**Step 1: Data Preprocessing**

We load the dataset, handle missing values, and encode categorical variables. Below is the code snippet used for this:



Dimple Patel-CT20244439775 4

**SafeCredit: AI-Based Loan Risk Prediction**

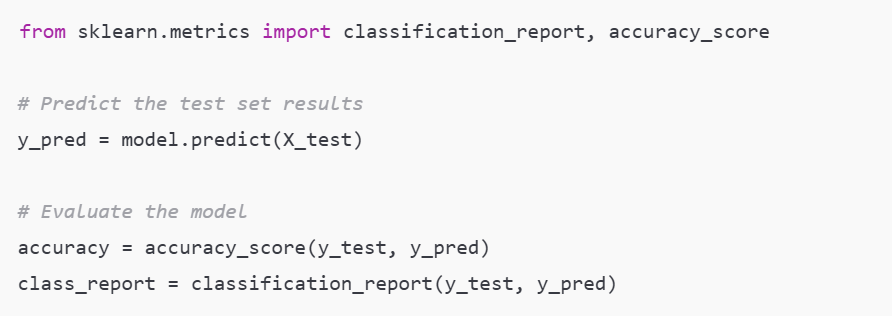
**Step 2: Model Selection**

We use the **Random Forest Classifier** for classification:



**Step 3: Model Evaluation**

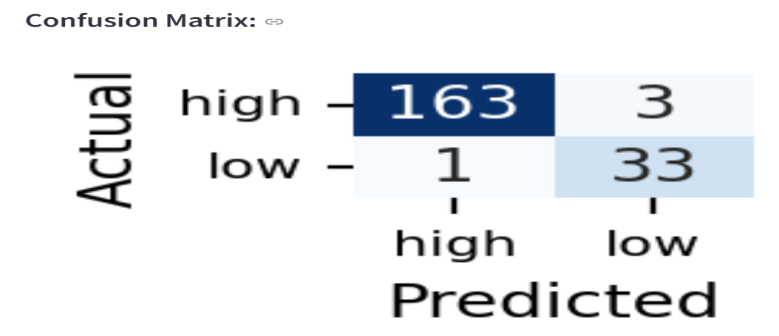
We evaluate the model's performance using accuracy and other metrics like precision, recall, and F1-score:



#### ****4.3 Hyperparameter Tuning****

Hyperparameters such as the number of trees, maximum depth, and minimum samples per leaf were tuned using **Grid Search** for optimal performance.

### ****4.4 Model Output and Visualization****



**Figure 1: Model Evaluation - Confusion Matrix**

Dimple Patel-CT20244439775 5

**SafeCredit: AI-Based Loan Risk Prediction**

Chapter 5

Standards Adopted

### ****5.1 Design Standards****

The model and interface were designed to be modular and scalable, with reusable functions for data preprocessing, model training, and prediction.

### ****5.2 Coding Standards****

Code follows **PEP 8** Python standards and includes proper documentation, making the code readable and maintainable.

### ****5.3 Testing Standards****

Testing was conducted using both unit tests for individual components and integration tests to ensure that the model and Streamlit interface work seamlessly together.

### ****5.4 Planning and Project Management****

The project was broken into phases:

Data collection and preprocessing

Model training and validation

UI development with Streamlit

Evaluation and reporting

Dimple Patel-CT20244439775 6

**SafeCredit: AI-Based Loan Risk Prediction**

Chapter 6

Conclusion and Future Scope

### ****6.1 Conclusion****

The project successfully developed a machine learning-based system to predict credit risk for loan applicants. The Random Forest model achieved a high level of accuracy, and the **Streamlit** interface provides a user-friendly way to interact with the model in real-time.

### ****6.2 Future Scope****

Future improvements could include:

Incorporating more advanced machine learning algorithms like **Gradient Boosting** or **Neural Networks** for better performance.

Expanding the features used in the model (e.g., including external data like social media activity, past loan history).

Improving the user interface for better visualization and data presentation.

Dimple Patel-CT20244439775 7

**A PROJECT REPORT**

**On**

**“SafeCredit: AI-Based Loan Risk Prediction”**

**Submitted to**

**Tata Consultancy Services**

**Submitted by**

**Dimple Patel**

**CT20244439775**