Enhancing Credit Risk Assessment through Advanced Machine Learning Techniques

CAPSTONE PROJECT PHASE-II

Phase – II Review 1 Report

Submitted by Group No 244

Sl. No.	Register Number	Name
1.	21BCE10014	Aryan Rai
2.	21BCE10443	Harshit Varshney
3.	21BCE11188	Adarsh Kanungo
4.	21BCE10470	Giya Ambasta
5.	21BCE10503	Saumadipta Chatterjee

in partial fulfillment of the requirements for the degree of

Bachlore of Engineering and Technology

VIT Bhopal University
Bhopal
Madhyapradhesh

January, 2025

Bonafide Certificate

Certified that this project report titled **“Enhancing Credit Risk Assessment through Advanced Machine Learning Techniques”** is the bonafide work of Group No 244 “21BCE10014 Aryan Rai, 21BCE10443 Harshit Varshney, 21BCE11188 Adarsh Kanungo, 21BCE10470 Giya Ambasta, 21BCE10503 Saumadipta Chatterjee**”** who carried out the project work under my supervision.

This project report (DSN4095-Capstone Project Phase-II) is submitted for the Project Viva-Voce examination held on 8th, January, 2025

Supervisor
Dr. Preetam Suman

Reviewer 1
Dr. Daood Saleem

Reviewer 2
Dr Sandeep Monga

Contents

CHAPTER 1: INTRODUCTION

The financial sector has undergone a massive transformation with the integration of technology, and credit risk assessment has emerged as a critical area for leveraging machine learning. Traditional methods relied on static rule-based systems and human intervention, often leading to inefficiencies and increased risks. As data availability grows, machine learning techniques provide opportunities for more accurate, automated, and scalable solutions.

In this project, we aim to develop a data-driven credit risk prediction model by analyzing customer data and external credit reports. By leveraging advanced machine learning techniques, the project seeks to provide financial institutions with a tool to make informed lending decisions, reduce non-performing assets (NPAs), and improve operational efficiency.

This report documents the progression of the project, from data preparation and exploratory analysis to feature engineering, model building, and evaluation. The following chapters outline the methodology, results, and learnings gained from this endeavor, showcasing the potential of machine learning in solving real-world financial challenges.

1.1 Motivation

The banking industry plays a crucial role in economic stability by offering financial services such as loans, credit, and investments. However, with an increase in loan demand, managing credit risk has become more challenging. Non-Performing Assets (NPA) are a significant threat to a bank's profitability, often resulting in financial losses and liquidity issues. As more individuals and businesses seek loans, banks need a more reliable method to evaluate the creditworthiness of applicants to minimize defaults and ensure sustainable lending practices.

Traditional credit assessment methods, while effective to an extent, rely heavily on manual processes and subjective judgment, which may lead to inconsistent decisions. Additionally, these methods are often slow and inefficient, unable to keep pace with the high volume of loan applications and the complex financial behaviour of customers.

This project is motivated by the need to develop an efficient, accurate, and scalable system that utilizes machine learning to predict loan defaults. By leveraging internal bank data and external sources such as CIBIL, we aim to build a model that not only streamlines the decision-making process but also enhances the accuracy of predicting whether a loan should be approved or denied. This solution could significantly reduce the risk of NPAs, helping banks improve their financial health while maintaining a responsible lending policy.

1.2 Objective

The primary objective of this project is to develop a machine learning-based credit risk model that can accurately predict the likelihood of a loan default. By analysing both internal bank data and external credit bureau data (CIBIL), the model will help banks in making informed lending decisions. This solution aims to streamline the loan approval process by reducing manual effort, improving prediction accuracy, and minimizing credit risk.

Specifically, the objectives of the project are as follows:

1. **Data Integration**: Combine internal banking datasets with external credit scores and related information to create a comprehensive dataset for analysis.
2. **Data Cleaning and Feature Engineering**: Handle missing data, outliers, and feature transformation to prepare the data for modeling. This includes removing null values and reducing multicollinearity among features.
3. **Model Development**: Build and train machine learning models using classification algorithms to predict the likelihood of loan approval or rejection based on various financial and demographic factors.
4. **Model Evaluation**: Evaluate the performance of the model using metrics like accuracy, precision, recall, and F1-score to ensure that it meets the required performance standards.
5. **Deployment and Integration**: Create a deployable version of the model for real-time credit decision-making in the bank's loan processing system.
6. **Reduction of NPAs**: Ultimately, the model seeks to reduce the number of non-performing assets (NPAs) by identifying high-risk applicants before loans are granted.

These objectives are aimed at enhancing the bank's ability to mitigate credit risk, improve profitability, and ensure more efficient loan processing.

CHAPTER 2: EXISTING WORK / LITERATURE REVIEW

Credit risk modeling is a crucial component of the financial industry, particularly for banking institutions that rely on accurate assessments of borrower creditworthiness to minimize the risk of loan defaults. Over the years, significant research and development have been conducted in this field, evolving from traditional statistical models to sophisticated machine learning techniques. This chapter reviews the existing work related to credit risk modeling, highlighting various methods, approaches, and their impact on the banking sector.

2.1 Traditional Approaches to Credit Risk Modeling

Historically, banks have relied on traditional statistical techniques, such as logistic regression and linear discriminant analysis (LDA), to predict credit risk. These models use a limited set of variables such as income, credit history, employment status, and debt-to-income ratios to assess whether a borrower is likely to default.

- **Credit Scoring Systems:** Credit scoring systems, such as FICO and CIBIL, have been widely used to evaluate an individual's creditworthiness. These scores are derived from an individual’s financial history, including payment behavior, outstanding debts, and credit inquiries. The scores categorize borrowers into risk levels, allowing banks to decide whether to approve or deny credit. While these systems are effective in evaluating basic risk, they lack the ability to analyze complex interactions between features.
- **Logistic Regression**: One of the earliest and most common methods used for binary classification in credit risk modeling is logistic regression. This technique models the probability of default by evaluating various predictor variables. Although logistic regression is easy to interpret and implement, it lacks the ability to capture complex relationships between variables.
- **Linear Discriminant Analysis (LDA)**: LDA is another statistical technique widely used for credit scoring, which classifies applicants into categories of "default" or "non-default" by creating a linear combination of predictor variables. While effective for well-separated classes, LDA struggles with non-linear relationships between the features.
- **Discriminant Analysis:** Discriminant analysis has been employed for classifying borrowers into default and non-default groups. Similar to logistic regression, it works well for smaller datasets but struggles with high-dimensional data and assumes normality of predictor variables, which may not hold true in real-world scenarios.

These traditional methods have the advantage of being simple and interpretable but may not perform well with large, complex datasets, which has led to the adoption of more advanced machine learning techniques.

2.2 Machine Learning Approaches

With the advent of big data and improved computational power, machine learning models have gained prominence in credit risk prediction. These models, including decision trees, random forests, gradient boosting machines, and deep learning networks, offer better accuracy and flexibility in dealing with complex datasets. Some of the most notable approaches are:

- **Decision Trees:** Decision trees are simple yet powerful algorithms that split data into branches based on feature values. They are easy to interpret and provide insights into the decision-making process. However, standalone decision trees are prone to overfitting, which limits their generalizability.
- **Random Forest:** Random Forest, an ensemble method, builds multiple decision trees and aggregates their predictions to improve accuracy and robustness. It reduces overfitting by introducing randomness in feature selection and data sampling. Studies have shown that Random Forest performs well in predicting default probabilities.
- **Gradient Boosting Machines (GBM)**: GBM models like XGBoost and LightGBM have become popular in recent years for their high predictive power. These models build on decision trees by iteratively improving predictions based on previous errors. They are known for their flexibility and efficiency in handling large, imbalanced datasets commonly found in banking.
- **Support Vector Machines (SVM)**: SVMs are another class of models that can separate data points using hyperplanes. Although effective for certain types of data, they require extensive feature scaling and can be computationally intensive for large datasets.
- **Deep Learning**: In more recent work, neural networks and deep learning models have been explored for credit risk modeling. These models can automatically extract features from the data and learn complex patterns, making them powerful for large datasets with intricate relationships between variables. However, they can be difficult to interpret, which is a significant concern in the banking sector where transparency is required.
- **Neural Networks:** Neural networks have been explored for credit risk modeling due to their ability to capture nonlinear relationships. Deep learning models, though less interpretable, have shown promise in analyzing unstructured data such as transaction histories and customer reviews.

2.3 Integration of External Data Sources

In addition to internal banking data, external data sources such as credit bureau information (e.g., CIBIL) and alternative data from social media, utility bills, and telecom data have become popular in recent years. These additional data points offer a more comprehensive view of a borrower's creditworthiness.

- **Credit Bureau Scores**: External credit scores like the CIBIL score in India are frequently used alongside internal data for credit risk assessment. These scores provide a standardized view of the borrower’s credit history, payment behavior, and outstanding loans, significantly improving predictive power.
- **Alternative Data**: The use of alternative data in credit scoring is gaining momentum. This includes information like mobile phone usage, online transaction data, and social media activity. While not yet as widespread, alternative data has been shown to improve predictions for applicants with limited traditional credit histories (commonly known as "thin-file" customers).

2.4 Challenges in Credit Risk Modeling

Despite the advancements in machine learning, there are still several challenges in credit risk modeling:

- **Data Quality**: Missing, incomplete, or incorrect data can skew the results of models, leading to inaccurate predictions. Handling missing values, such as the -99999 placeholder in our dataset, is critical for ensuring model reliability.
- **Imbalanced Data**: Credit risk datasets are often highly imbalanced, with far fewer defaults than non-defaults. This can lead to biased models that are more likely to predict non-defaults, as the algorithm tends to favor the majority class. Techniques such as oversampling, undersampling, or using specific evaluation metrics like AUC-ROC are commonly used to address this issue.
- **Interpretability vs. Complexity**: While machine learning models like random forests and gradient boosting machines offer higher accuracy, they often lack interpretability. In the banking sector, regulatory bodies require transparency in how credit decisions are made. Balancing accuracy and interpretability remains a key challenge in the deployment of these models.

2.5 Gaps Addressed by our Project

While existing studies have successfully applied machine learning techniques to credit risk modeling, several gaps remain:

1. **Feature Engineering:** Many studies overlook the importance of feature selection and engineering. Our project employs statistical methods such as Chi-Square tests, Variance Inflation Factor (VIF), and ANOVA to refine input features, ensuring higher model interpretability and accuracy.
2. **Integration of Internal and External Data:** Few models leverage both internal bank data and external credit bureau data. By combining these datasets, our approach provides a more comprehensive view of borrower risk.
3. **Class Imbalance Handling:** Imbalanced datasets are a common challenge in credit risk modeling. Our project uses techniques such as ensemble methods and metrics evaluation (e.g., precision, recall, F1-score) to ensure balanced performance across all classes.
4. **Multi-Class Classification:** While most research focuses on binary classification (default vs. non-default), our project extends the scope to multi-class classification, predicting loan approval priorities (P1, P2, P3, P4).

2.6 Comparison of existing approaches and our model

Feature	Traditional Methods	Machine Learning (Existing)	Our Project
Data Handling	Small, structured datasets	Large datasets	Combined internal and external data
Feature Selection	Manual	Automated	Statistical techniques (VIF, ANOVA)
Modeling Approach	Logistic Regression	Ensemble Models (e.g., RF)	Random Forest, XGBoost, Decision Tree
Performance on Imbalanced Data	Poor	Moderate	Addressed with advanced techniques
Classification Type	Binary	Binary	Multi-class

2.7 Conclusion

The field of credit risk modeling has evolved significantly, moving from traditional statistical methods to more sophisticated machine learning algorithms. With the integration of external credit bureau data and alternative data sources, modern credit risk models offer greater predictive power and flexibility. However, challenges such as data quality, model interpretability, and handling imbalanced datasets remain. This project builds on these existing methods, aiming to develop a machine learning-based credit risk model that addresses these challenges while delivering accurate, actionable predictions for loan approval decisions.

CHAPTER 3: FRONT END, BACKEND, AND SYSTEM REQUIREMENT

3.1 Front End & Back End (Technology Stack Used)

In this project, a machine learning-based credit risk model is developed to predict loan approval decisions. The technology stack used in this project involves the following components:

Front End: The front-end serves as the interface for users to interact with the model. For this project, the front end could be either a **web-based interface** or a **desktop application**, depending on the deployment method.

- **Web Interface**: If deployed on the cloud, the front end can be built using **HTML5**, **CSS3**, and **JavaScript** for a responsive user interface. Libraries such as **React.js** or **Vue.js** could be used for building dynamic and interactive components. This allows users to input applicant data, receive credit approval predictions, and view model results in real-time.
- **Desktop Application**: In case of a desktop deployment, a **Python-based executable (EXE)** can be created using libraries such as **PyQt** or **Tkinter** for designing the user interface. This option provides a standalone tool for bank officers to use the model on their computers without requiring internet connectivity.

Back End: The back-end is responsible for handling the logic, managing the data, and interacting with the machine learning model. The back end for this project involves the following components:

- **Python (Flask/Django)**: The core of the back-end will be written in Python, utilizing frameworks like **Flask** or **Django** to handle the logic and API requests. Flask is lightweight and suitable for small projects, while Django offers a more comprehensive framework for larger systems.
- **Machine Learning Libraries**: Libraries like **Scikit-learn**, **XGBoost**, **LightGBM**, and **Pandas** are used for training and deploying the machine learning model. **Pickle** or **Joblib** will be used for model serialization, allowing the trained model to be saved and loaded for predictions.

Deployment Options:

- **Cloud Deployment**: The application can be deployed on cloud platforms like **AWS** or **Microsoft Azure** using services such as **AWS Lambda** or **Azure App Service** for model hosting. This will enable real-time predictions, making the system scalable and accessible from any location.
- **Executable Deployment**: Alternatively, an **executable EXE** file can be generated using **PyInstaller** or **cx_Freeze**. This option is ideal if the system is intended for internal use within the bank, where no external connectivity is required.

3.2 System Requirements

Hardware Requirements:

- **Processor**: Minimum 2.4 GHz, multi-core processor (10th Gen Intel i5 or higher recommended).
- **RAM**: At least 8 GB of RAM is required for efficient model training and deployment.
- **Storage**: A minimum of 10 GB of free disk space is required to store the datasets, libraries, and model files.
- **GPU**: A dedicated Graphics Processing Unit (GPU) like **NVIDIA Tesla** or **GeForce RTX** is optional but recommended for faster training, especially for complex models or deep learning applications.

Software Requirements:

- **Operating System**: Windows 10/11, Linux (Ubuntu), or macOS.
- **Python Version**: Python 3.8 or above.
- **Machine Learning Libraries**: Scikit-learn, XGBoost, LightGBM, Pandas, NumPy, and Matplotlib.
- **Development Tools**:
    - **Jupyter Notebook** or **Spyder** for writing and testing code.
    - **Anaconda** for managing packages and dependencies.
- **IDE**: Spyder 5.1.5 for development purposes.
- **Web Framework**: Flask or Django for creating the back end.
- **Database**: SQLite or MySQL for data storage and retrieval.
- **Deployment Tools**:
    - **AWS CLI** or **Azure CLI** for cloud deployment.
    - **PyInstaller** or **cx_Freeze** for packaging the application into an executable EXE for desktop deployment.

CHAPTER 4: METHODOLOGY

In this chapter, we discuss the design and architecture of the credit risk prediction system, the working principle behind the machine learning models used, and the results and analysis of the model's performance. This methodology demonstrates the step-by-step process of how the system was built and the key findings from its evaluation.

4.1 System Design / Architecture

The system architecture for the credit risk prediction model is designed with scalability, flexibility, and efficiency in mind. The model predicts whether a customer's loan application should be approved based on features extracted from internal bank data (Case Study 1) and external CIBIL data (Case Study 2).

The system is structured into the following components:

1. Data Collection:
The system gathers data from both internal sources (bank's product holdings and loan data) and external credit bureau reports (CIBIL scores and reports). This raw data is stored in a secure database (such as MySQL or SQLite) for further processing.

2. Data Preprocessing and Feature Engineering:
This stage is crucial for transforming the raw data into a format suitable for machine learning. Various steps, such as handling missing values (like replacing -99999 values), normalizing numerical features, encoding categorical variables, and combining the datasets on `PROSPECTID`, are applied. Additionally, exploratory data analysis (EDA) is performed to understand patterns in the data.

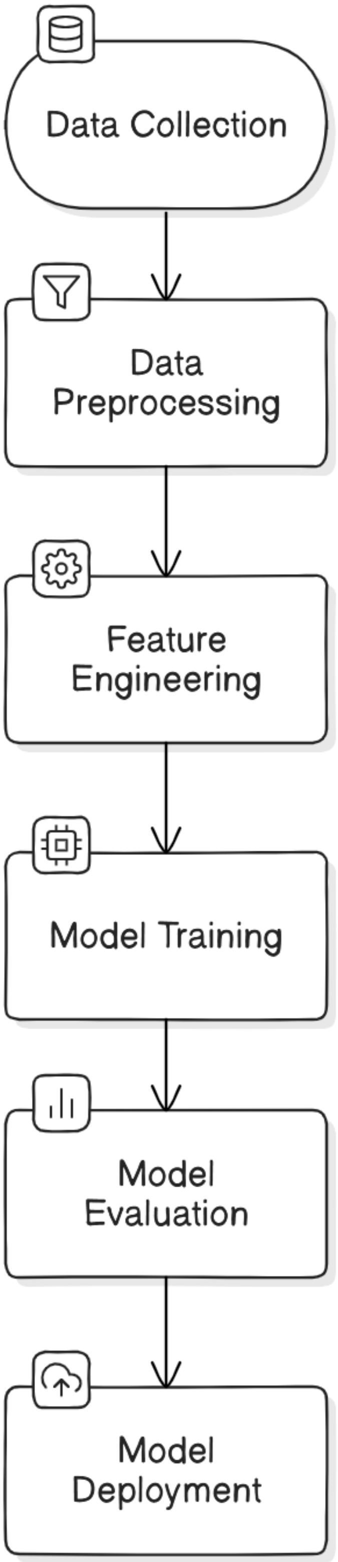
3. Model Training:
The pre-processed data is used to train machine learning models. Various algorithms such as **Logistic Regression**, **Random Forest**, and **XGBoost** are employed. Model training involves splitting the data into training and testing sets, using cross-validation to prevent overfitting, and tuning hyperparameters to improve model performance.

4. Model Evaluation:
Once the models are trained, they are evaluated based on their predictive performance using metrics like **Accuracy**, **Precision**, **Recall**, **F1-Score**, and **ROC-AUC Score**. Feature importance analysis is conducted to determine the most influential factors contributing to loan approval or denial.

5. Model Deployment:
The model is deployed on a cloud platform (like AWS or Azure) or converted into an executable (EXE) file for local deployment. The system is designed to take user input (e.g., customer information), run the prediction model, and return the result (approve/deny) in real-time.

The architecture of the system is visualized in the diagram below.

System Architecture Diagram:



4.2 Working Principle

The working principle of the project is divided into multiple stages:

4.2.1 Data Collection and Integration

1. **Datasets Used:**
   - **Internal Product Dataset (Case Study 1):** Contains details of customer loan product holdings.
   - **External CIBIL Dataset (Case Study 2):** Provides credit scores and delinquency information.
2. **Integration:** The datasets are merged using the unique identifier PROSPECTID, ensuring a comprehensive view of customer behavior.

4.2.2 Data Preprocessing

1. **Handling Missing Values:** Null values represented as –99999 are identified and treated based on their frequency.
2. **Outlier Detection and Removal:** Univariate and multivariate analysis techniques are used to remove extreme values.
3. **Data Standardization:** Ensures uniform scaling of features for better model performance.

4.2.3 Feature Engineering

1. **Statistical Methods:**
   - **ANOVA Test:** Identifies significant numerical features correlated with the target variable.
   - **Chi-Square Test:** Assesses the association between categorical features and the target.
   - **Variance Inflation Factor (VIF):** Removes multicollinearity among numerical features.
2. **Feature Encoding:**
   - Categorical features such as MARITALSTATUS and GENDER are converted using one-hot encoding.
   - Ordinal features like EDUCATION are encoded based on logical ordering (e.g., SSC < 12TH < GRADUATE).

4.2.4 Model Development

1. **Random Forest Classifier:**
   - An ensemble model that aggregates predictions from multiple decision trees.
   - Achieves an accuracy of 76% on the test dataset.
2. **XGBoost:**
   - Optimized gradient boosting algorithm that excels in handling imbalanced datasets.
   - Achieves the highest accuracy of 78% with improved F1-scores for classes `P1`, `P2`, and `P4`.
3. **Decision Tree Classifier:**
   - A simple, interpretable model achieving an accuracy of 71%, serving as a baseline.

4.2.5 Model Evaluation Metrics

- **Accuracy:** Measures overall correctness of predictions.
- **Precision, Recall, F1-Score:** Evaluates class-wise performance, especially for imbalanced classes.
- **Confusion Matrix:** Visualizes model's prediction capabilities for each class.

4.3 Results and Discussion

4.3.1 Performance Metrics

The performance of each model is summarized in *Table 1* below:

Model	**Accuracy**	**Precision (P1)**	**Recall (P1)**	**F1-Score (P1)**	**Precision (P3)**	**Recall (P3)**	**F1-Score (P3)**
Random Forest	76.37%	0.837	0.704	0.765	0.442	0.211	0.286
XGBoost	78.00%	0.824	0.761	0.791	0.476	0.309	0.375
Decision Tree	71.00%	0.721	0.727	0.724	0.346	0.330	0.338

Table 1: Model Performance Metrics

4.3.2 Insights from Results

1. **Best Model:** XGBoost outperforms other models with the highest accuracy (78%) and balanced class-wise performance.

2. **Class Imbalance Impact:** The lower performance on class P3 highlights the need for oversampling or alternative class imbalance techniques.

3. **Feature Importance:** Features like CIBIL Score, Time Since Recent Payment, and Total Missed Payments were found to have the highest influence on predictions.

4.3.3 Visual Representation

1. **Confusion Matrix:** A confusion matrix for the XGBoost model is shown below:

Actual / Predicted	P1	P2	P3	P4
P1	1500	200	50	100
P2	100	1700	80	120
P3	50	70	250	60
P4	80	100	60	900

2. **Feature Importance Plot:** The top 5 most important features for XGBoost are:
    - CIBIL Score
    - Time Since Recent Payment
    - Total Missed Payments
    - Credit Card Utilization
    - Personal Loan Enquiries (Last 6 Months)

Individual Contribution by Members

The following section provides a detailed description of the contributions made by each team member throughout the project. Each member played a vital role in ensuring the success of the project, with responsibilities ranging from data handling to report writing and presentation preparation.

1) Aryan Rai (21BCE10014) - Feature Engineering, EDA, and Machine Learning Model Implementation (Main Report)

Aryan Rai played a pivotal role in preparing the dataset for machine learning by performing **Exploratory Data Analysis (EDA)** and **Feature Engineering** to enhance the predictive performance of the models. Key contributions include:

1. **Feature Engineering**:
    - Conducted thorough analysis to identify significant features using statistical techniques such as:
        - **Chi-Square Test**: To evaluate the dependency of categorical features on the target variable.
        - **ANOVA Test**: To analyze numerical features for their correlation with loan approval classes.
        - **Variance Inflation Factor (VIF)**: To remove multicollinearity among numerical features, ensuring independent predictors.
    - Encoded categorical variables using one-hot encoding and ordinal encoding (e.g., educational levels such as SSC, 12TH, GRADUATE).
    - Selected the most impactful features using systematic feature reduction techniques, narrowing the dataset to relevant variables.
2. **EDA (Exploratory Data Analysis)**:
    - Performed in-depth data visualization using Matplotlib and Seaborn to uncover trends and anomalies within the dataset.
    - Analyzed relationships between features such as "Total Missed Payments" and "CIBIL Scores" to understand their impact on credit risk.
    - Addressed missing values and outliers without compromising data integrity, adhering to financial industry standards.
3. **Assistance in Machine Learning Models**:
    - Collaborated in the implementation of models such as Random Forest, XGBoost, and Decision Tree.
    - Helped optimize hyperparameters to achieve better performance and interpretability.
    - Documented all processes and findings in the main project report, ensuring clarity and structure.

2) Harshit Varshney 21BCE10443 - Main Report and Machine Learning Model Selection & Implementation

Harshit Varshney contributed significantly to the **machine learning pipeline**, focusing on model selection, training, and performance comparison. Key contributions include:

1. **Model Selection**:
    - Evaluated multiple machine learning algorithms such as Random Forest, XGBoost, and Decision Tree for their suitability in multi-class classification.
    - Chose the optimal models based on accuracy, precision, recall, and F1-score metrics.
2. **Implementation**:
    - Trained, tested, and validated models using an 80-20 train-test split.
    - Used Scikit-learn and XGBoost libraries for implementation, ensuring robust and scalable solutions.
    - Focused on handling class imbalances (e.g., minority class P3) using ensemble techniques and performance adjustments.
3. **Comparative Analysis**:
    - Compared the models in terms of:
        - Overall accuracy: XGBoost achieved the highest accuracy (78%).
        - Class-wise metrics: Identified strengths and weaknesses of each model (e.g., lower F1-score for class P3 in Random Forest).
    - Documented the trade-offs between accuracy, interpretability, and computational cost for deployment planning.
4. **Report Contribution**:
    - Co-authored the main report with detailed sections on machine learning methodologies and results discussion.
    - Prepared visual aids, including confusion matrices and feature importance graphs, for better interpretation.

3) Adarsh Kanungo 21BCE11188 - Dataset Analysis and Preparation

Adarsh Kanungo ensured the dataset was clean, consistent, and ready for analysis by focusing on **data preparation and preprocessing**. Key contributions include:

1. **Handling Missing Values**:
    - Identified missing data patterns in both datasets (internal bank data and CIBIL data).
    - Removed rows where significant missing values were present (e.g., age of oldest trade line with -99999 as placeholder).
    - Dropped columns with over 10,000 missing values to maintain data integrity.
2. **Dataframe Merging**:
    - Merged internal and external datasets using the unique PROSPECTID column.
    - Ensured consistency and eliminated nulls during the merge process.
3. **Dataset Cleaning**:

    - Addressed outliers using statistical techniques.
    - Standardized numerical columns to ensure compatibility across features (e.g., normalizing "Time Since Recent Payment").
4. **Report Contribution**:
    - Documented the entire data preprocessing pipeline in the report.
    - Highlighted challenges faced during data cleaning and the solutions implemented.

4) Giya Ambasta 21BCE10470 - Research Study, Dataset Selection, and Report & PPT Support

Giya Ambasta contributed extensively to the **planning phase of the project** and ensured the datasets used were relevant and comprehensive. Key contributions include:

1. **Research Study**:
    - Conducted detailed research on credit risk assessment methodologies, providing the team with insights into traditional and modern approaches.
    - Studied financial concepts such as GNPA, NNPA, and DPD to align project objectives with banking standards.
2. **Dataset Selection**:
    - Analyzed multiple publicly available datasets and evaluated their relevance, completeness, and applicability to the project.
    - Chose the final datasets based on quality and their ability to support multi-class classification.
3. **Report and PPT Preparation**:
    - Assisted in writing key sections of the report, particularly the introduction and existing work.
    - Designed and structured the project presentation for Phase II Review.

5) Saumadipta Chatterjee 21BCE10503 - Research Study and PPT Contribution

Samyadeepta Chatterjee provided valuable insights during the research phase and supported the team in understanding the datasets and their use. Key contributions include:

1. **Research Study**:
    - Researched the use and interpretation of external credit bureau data (e.g., delinquency levels and CIBIL scores).
    - Studied how credit risk classification aligns with real-world banking practices, ensuring the project’s relevance.
2. **Dataset Interpretation**:
    - Helped the team understand the significance of key features such as "Missed Payments" and "Credit Card Utilization."
    - Provided feedback on the feature engineering process to improve the model’s effectiveness.
3. **PPT Contribution**:

- Assisted in preparing the Phase II Review presentation by creating slides summarizing research findings and model performance.
- Ensured the presentation was visually appealing and logically structured.

CHAPTER 5: CONCLUSION

The project aimed to design and implement a machine learning-based credit risk prediction system to assist banks in making informed lending decisions. By combining internal bank datasets and external credit bureau data, the system successfully leverages advanced data processing techniques, feature engineering, and machine learning algorithms to classify loan approval priorities.

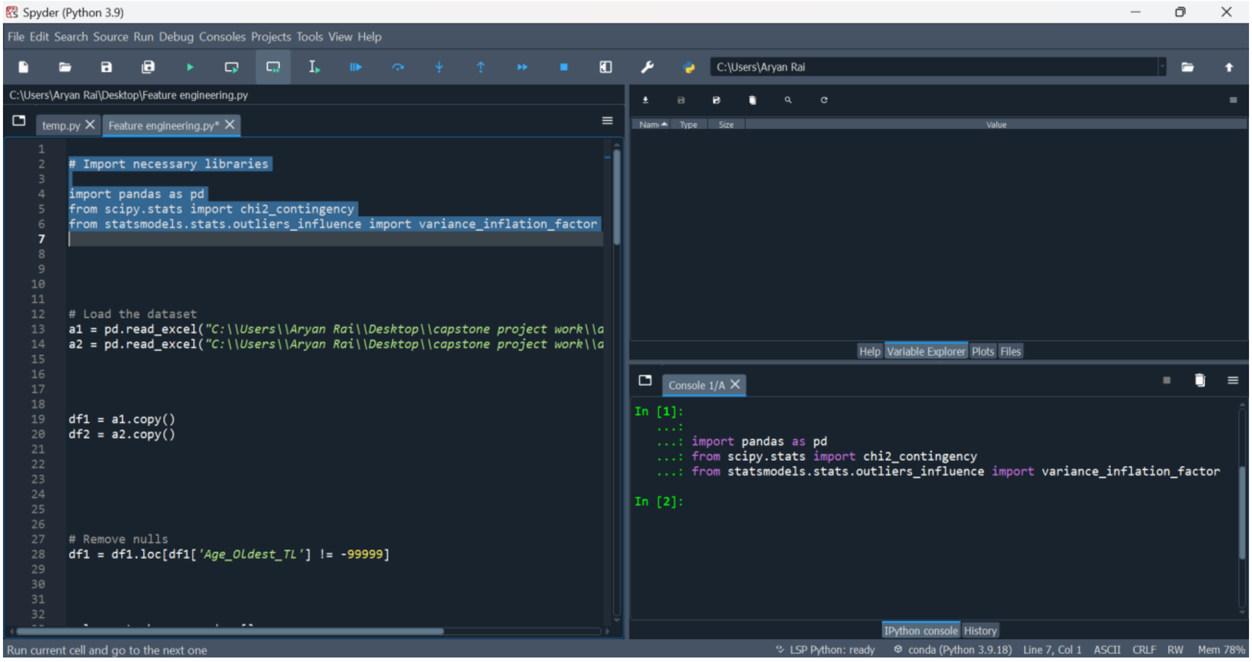
Code Workflow

The workflow of the code can be summarized in the following steps:

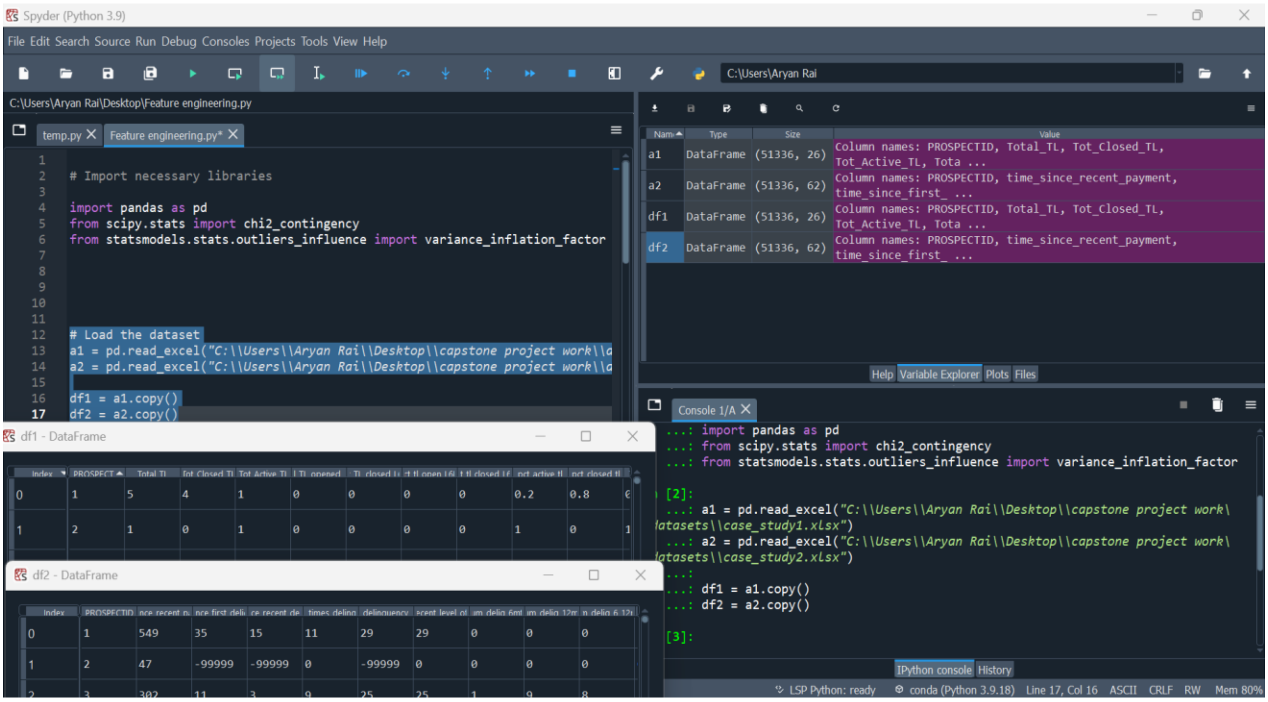
1. **Data Ingestion**:
    - Two datasets (internal and external) were merged using a unique identifier (PROSPECTID).
    - Missing values were identified and handled based on predefined rules (e.g., removing rows with –99999 values).
2. **Data Preprocessing**:
    - Outliers were removed to maintain data consistency.
    - Categorical features were encoded using techniques like one-hot encoding and ordinal encoding.
    - Numerical features were standardized, and multicollinearity was addressed using Variance Inflation Factor (VIF).
3. **Feature Selection**:
    - Statistical tests such as Chi-Square, ANOVA, and VIF were employed to identify significant features.
    - The final dataset consisted of features that were most relevant to predicting credit risk.
4. **Machine Learning Models**:
    - Three models (Random Forest, XGBoost, and Decision Tree) were implemented to predict loan approval categories (P1, P2, P3, P4).
    - Model evaluation metrics (accuracy, precision, recall, F1-score) were computed to compare their performance.
5. **Results Visualization**:
    - Confusion matrices and feature importance plots were generated to interpret model performance.
    - XGBoost emerged as the best-performing model with an accuracy of 78%.

5.1 Project Screenshot

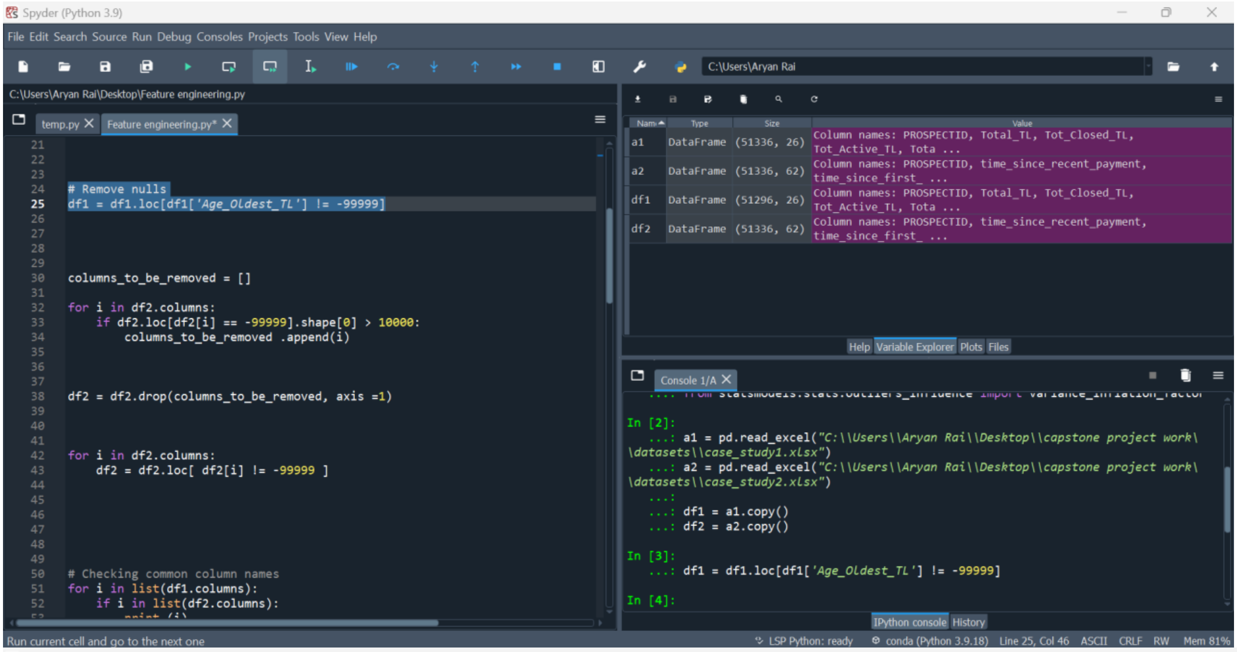
The first step is to load the necessary Python libraries that help with data handling, statistical tests, and multicollinearity checks.

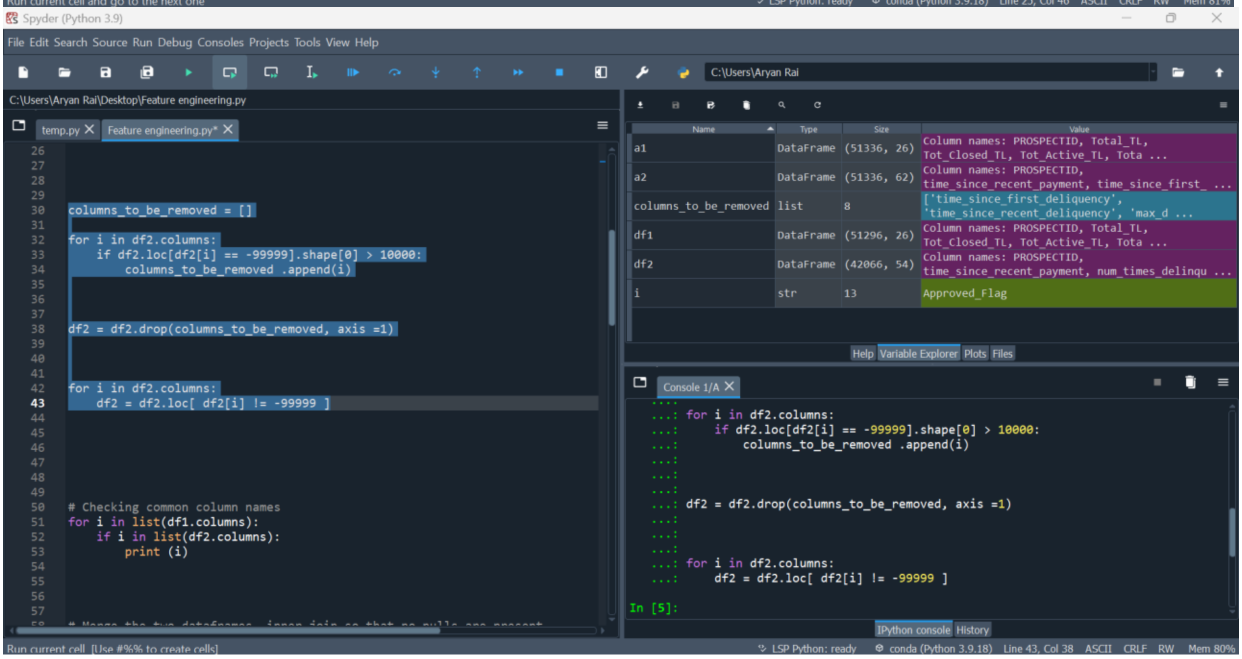


Two Excel files, `case_study1.xlsx` (internal product data) and `case_study2.xlsx` (CIBIL data), are loaded into Python and copied into dataframes `df1` and `df2` for further analysis.

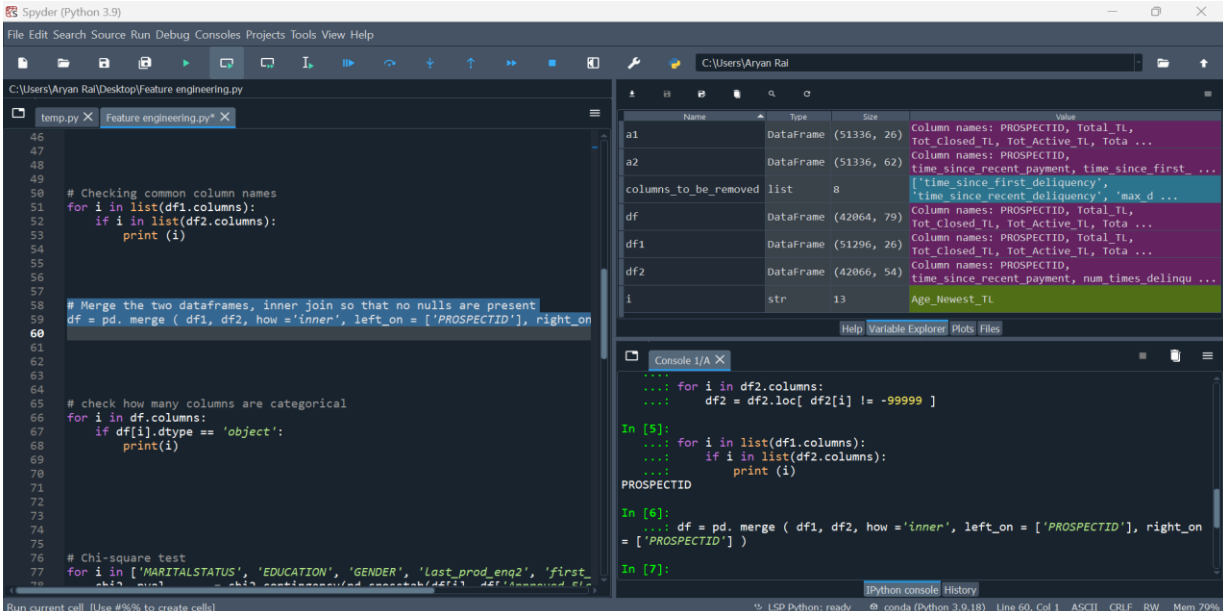


The code identifies missing values (represented as `-99999`) in both datasets. In `df1`, rows where the oldest account age is `-99999` are removed. In `df2`, columns with more than 10,000 missing values are dropped, and remaining rows with null values are also removed.





The two datasets are merged based on a common column `PROSPECTID`. This ensures that we have a unified dataset with no missing values, ready for analysis.



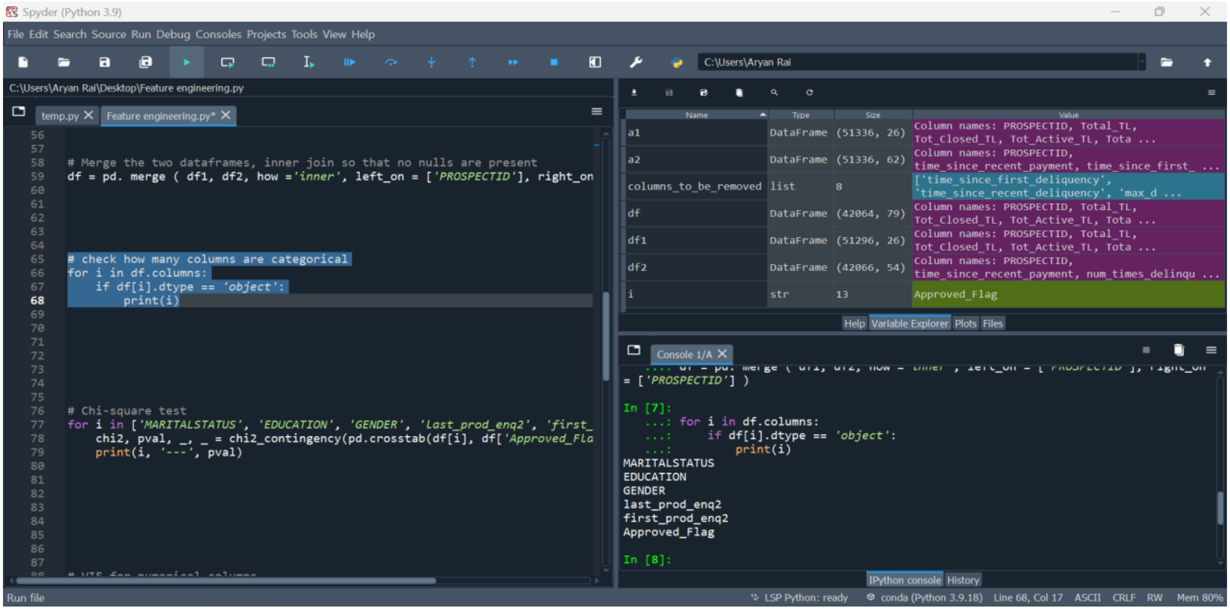
Finally, we have our single Dataset Ready for Modeling

df - DataFrame

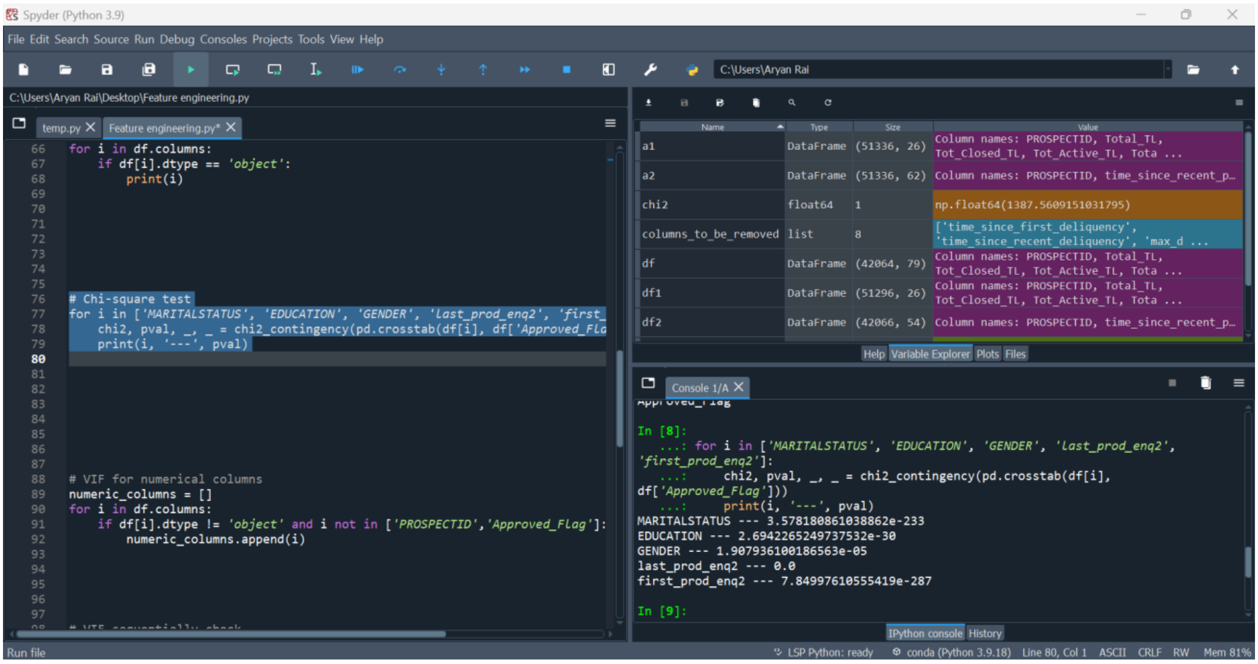
Index	PROSPECTID	Total TL	Tot Closed TL	Tot Active TL	TL opened	TL closed	t tl open L6	t tl closed L6	pct active tl	pct closed tl	TL opened L	TL closed L1	t tl open L12	tl closed L1	t Missed Pm	Auto TL	CC TL	Consumer TL	Gold TL	Home TL	PL TL
0	1	5	4	1	0	0	0	0	0.2	0.8	0	0	0	0	0	0	0	0	1	0	4
1	2	1	0	1	0	0	0	0	1	0	1	0	1	0	0	0	0	1	0	0	0
2	3	8	0	8	1	0	0.125	0	1	0	2	0	0.25	0	1	1	0	6	1	0	0
3	5	3	2	1	0	0	0	0	0.333	0.667	0	0	0	0	0	1	0	0	0	0	0
4	6	6	5	1	0	0	0	0	0.167	0.833	0	1	0	0.167	0	4	0	0	2	0	0
5	8	6	4	2	0	0	0	0	0.333	0.667	1	2	0.167	0.333	0	1	0	0	0	0	0
6	9	1	0	1	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0
7	11	7	2	5	1	0	0.143	0	0.714	0.286	3	1	0.429	0.143	0	2	1	2	0	0	2
8	13	2	2	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0
9	14	2	1	1	1	0	0.5	0	0.5	0.5	1	1	0.5	0.5	0	0	0	2	0	0	0
10	15	1	1	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0
11	16	8	1	7	7	1	0.875	0.125	0.875	0.125	8	1	1	0.125	2	1	1	5	0	0	0
12	17	1	0	1	0	0	0	0	1	0	1	0	1	0	0	0	0	0	0	0	0
13	20	7	5	2	1	0	0.143	0	0.286	0.714	1	1	0.143	0.143	0	0	0	0	2	0	0
14	21	1	0	1	0	0	0	0	1	0	1	0	1	0	0	0	0	1	0	0	0
15	22	5	2	3	0	1	0	0.2	0.6	0.4	2	1	0.4	0.2	1	3	0	1	0	0	1
16	23	1	1	0	0	1	0	1	0	1	0	1	0	1	0	1	0	0	0	0	0
17	25	3	1	2	0	0	0	0	0.667	0.333	1	1	0.333	0.333	0	0	0	0	0	0	0
18	27	1	1	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0

Format Resize Background color Column min/max Save and Close Close

The code then checks which columns contain categorical data (e.g., "Marital Status"). This is important for statistical tests later.

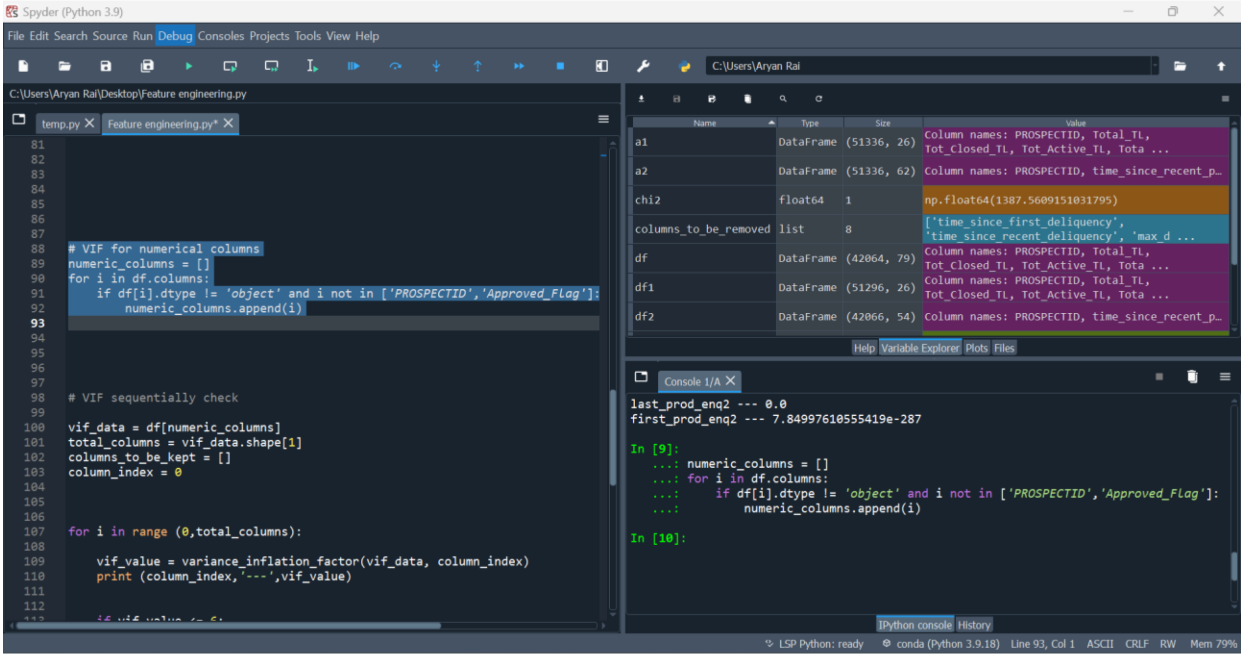


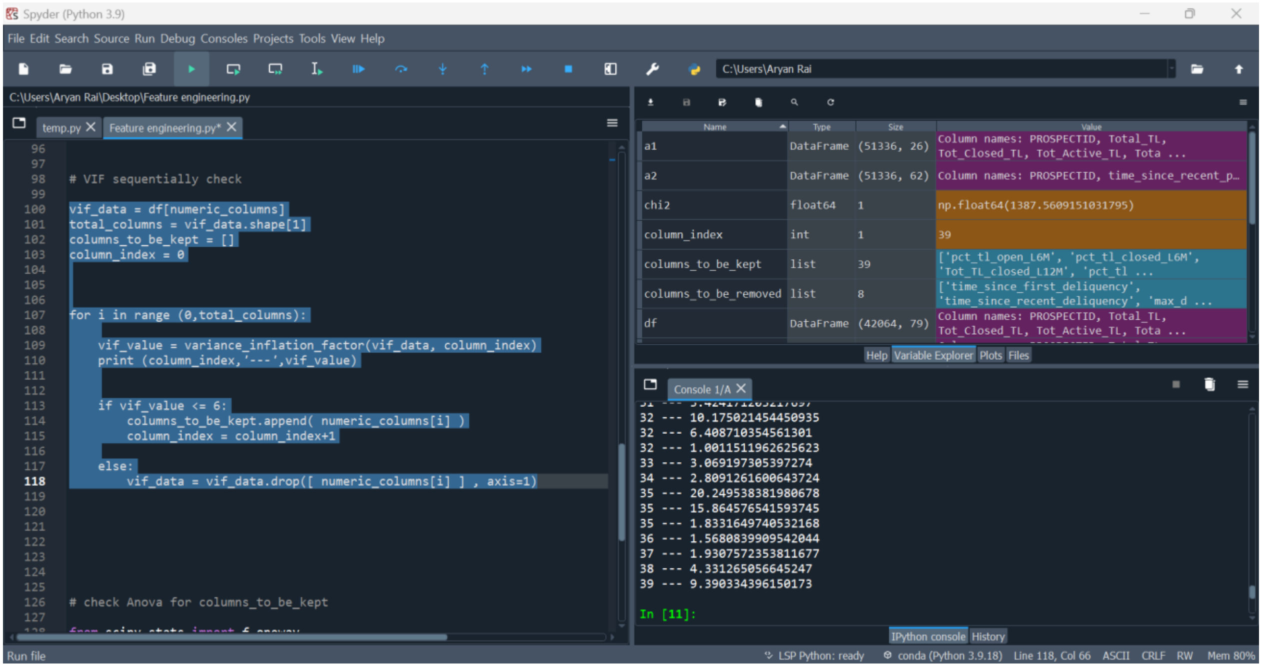
Chi-Square Test for Categorical Features:
For each categorical feature (like Marital Status, Education, Gender, etc.), a chi-square test is performed to check if it is associated with the loan approval flag (`Approved_Flag`). This test helps determine which categorical features have a significant relationship with the target variable.



Checking Multicollinearity with VIF:

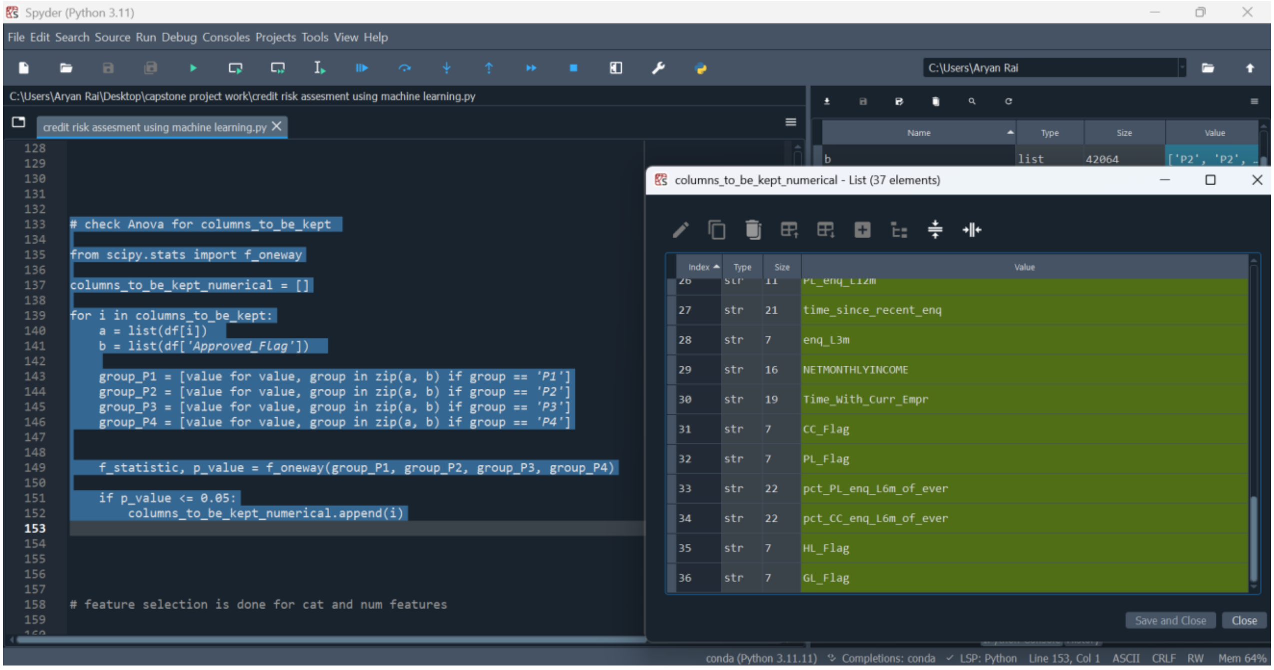
For numerical columns, multicollinearity is checked using the **Variance Inflation Factor (VIF)**. Multicollinearity occurs when two or more features are highly correlated, and removing such features can improve the model's accuracy. The code removes any features with a VIF value greater than 6.



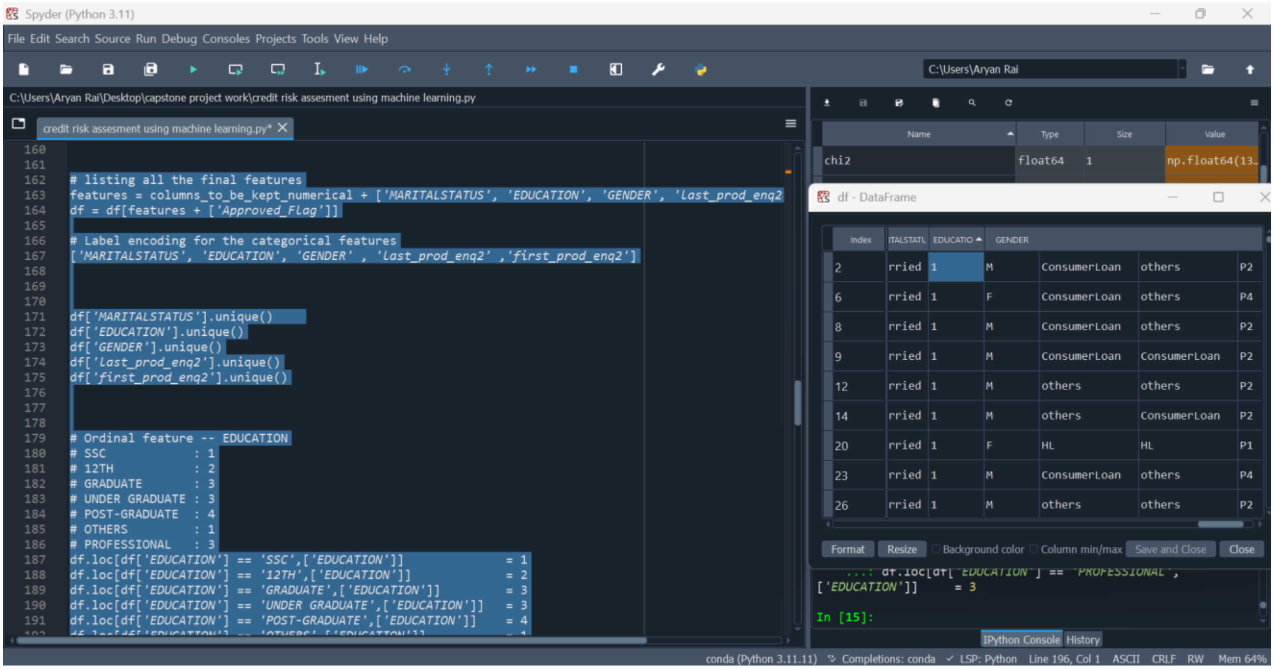


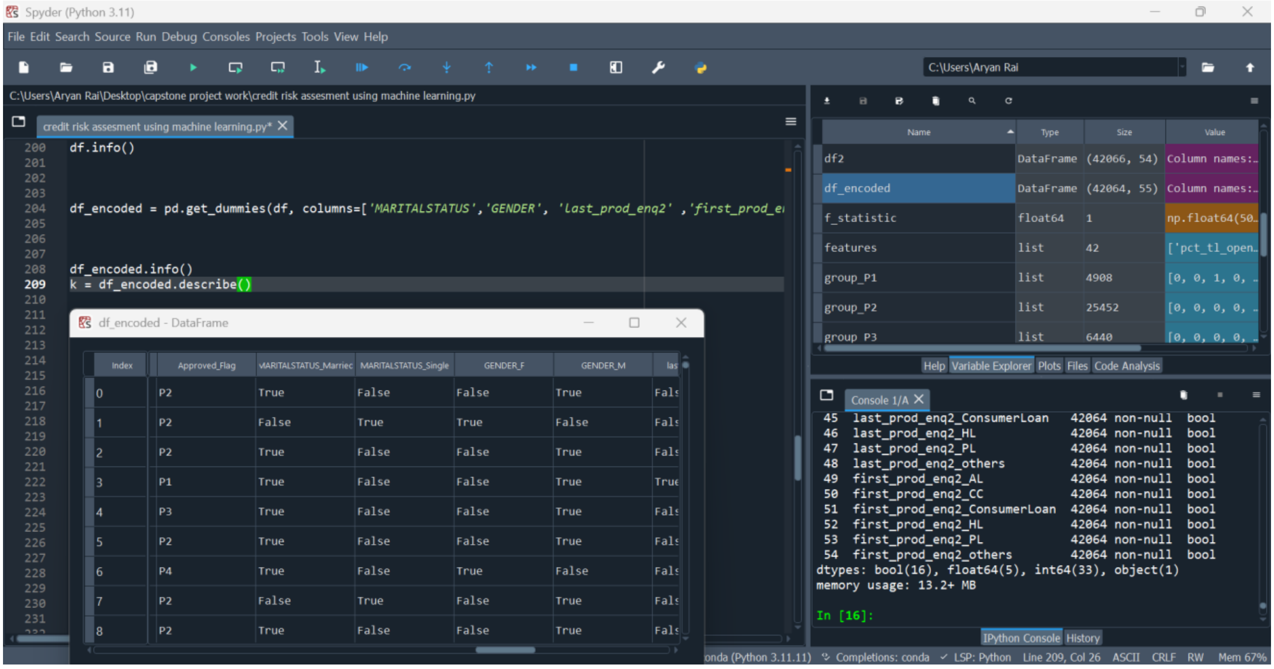
ANOVA Test for Numerical Features:

Finally, the code performs an **ANOVA** test on numerical features to check whether they have a significant relationship with the target variable. Only features with a p-value less than or equal to 0.05 are kept for further analysis.

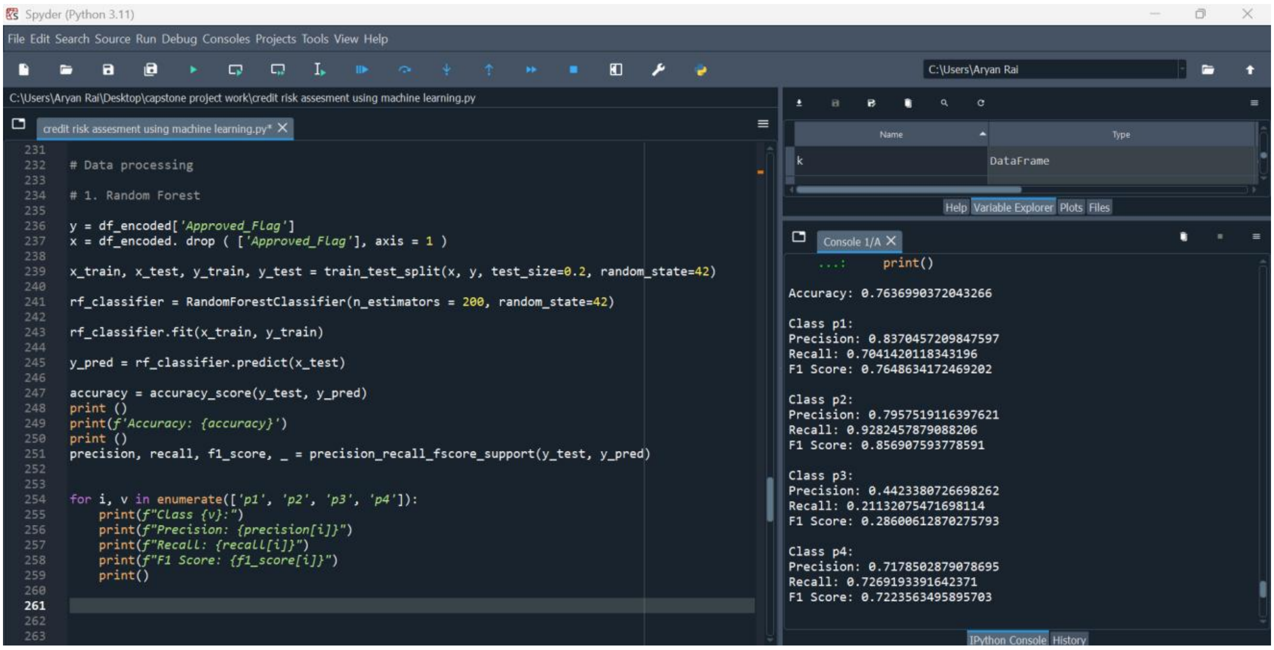


Encoding Categorical Features: Ordinal encoding is applied to the EDUCATION feature because it can be logically encoded. One-hot encoding is applied to categorical variables like MARITALSTATUS, GENDER, and others because it cannot be logically encoded for machine learning model implementation.

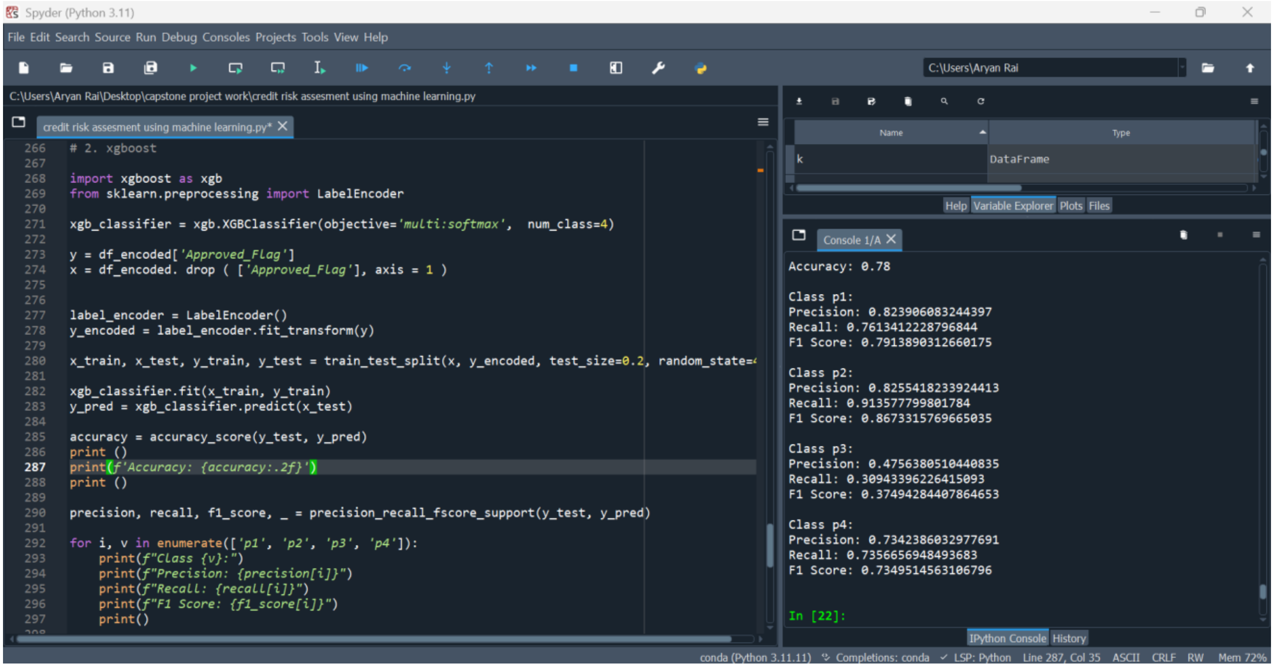




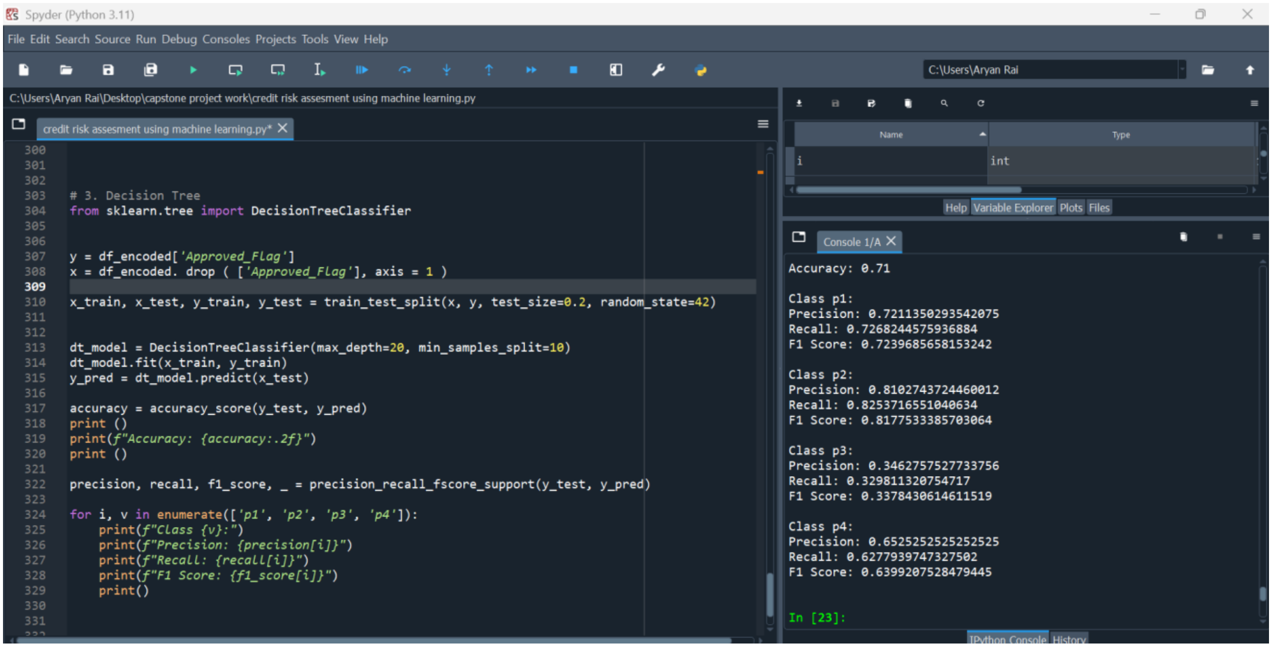
Applied Random Forest Model - The Random Forest model demonstrates a strong performance with an accuracy of ~76%.



Applied Xgboost model - XGBoost outperforms Random Forest in terms of overall accuracy (78% vs. 76%).



Applied Decision Tree - The Decision Tree is less accurate compared to XGBoost and Random Forest.



5.2 Summary:

This project highlights the potential of machine learning in transforming credit risk assessment by offering data-driven insights and efficient classification methods. By integrating statistical feature engineering techniques and robust machine learning models, the system ensures accuracy, scalability, and relevance to banking operations. Future work can focus on deploying the model in a live environment and incorporating additional data sources for enhanced predictions.

References and Publications

1. M. Sudhakar and C.V.K. Reddy, "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), vol. 5(3), pp. 705-718, 2016.
2. J. H. Aboobyda and M.A. Tarig, "Developing Prediction Model of Loan Risk in Banks Using Data Mining," Machine Learning and Applications: An International Journal (MLAIJ), vol. 3(1), pp. 1–9, 2016.
3. K. Kavitha, "Clustering Loan Applicants based on Risk Percentage using K-Means Clustering Techniques," International Journal of Advanced Research in Computer Science and Software Engineering, vol. 6(2), pp. 162–166, 2016.
4. Trishita Saha, Saroj Kumar Biswas, Saptarsi Sanyal, Binoti Patro, Biswajit Purkayastha, “Credit Risk Prediction Using Machine Learning Analytics: An Ensemble Model,” *International Journal of Management and Applied Science*, Vol. 9, Issue 9, 2023.
5. Sudhamathy G., "Credit Risk Analysis and Prediction Modelling of Bank Loans Using R," *International Journal of Engineering and Technology (IJET)*, Vol. 8, No. 5, Oct-Nov 2016, DOI: 10.21817/ijet/2016/v8i5/160805414.
6. S. Dahiya, S.S. Handa, N.P. Singh, "A Feature Selection Enabled Hybrid-Bagging Algorithm for Credit Risk Evaluation," *Expert Systems*, 2017. DOI: 10.1111/exsy.12217.
7. A.A. Turjo, K.O. Alabi, S.O. Abdulsalam, R.O. Ogundokun, "Credit Risk Assessment Model by Analyzing Different Machine Learning Algorithms," *4th International Conference on Information and Communications Technology (ICOIACT)*, 2021. DOI: 10.1109/ICOIACT53268.2021.9563995.
8. Y-C Chang, K-H Chang, G-J Wu, "Application of eXtreme Gradient Boosting Trees in the Construction of Credit Risk Assessment Models for Financial Institutions," *Applied Soft Computing Journal*, 2018. DOI: 10.1016/j.asoc.2018.09.029.
9. Y. Zhang, H. Jia, Y. Diao, M. Hai, H. Li, "Research on Credit Scoring by Fusing Social Media Information in Online Peer-to-Peer Lending," *Procedia Computer Science*, 2016. DOI: 10.1016/j.procs.2016.07.055.
10. V. Kumar L, Natarajan S, Keerthana S, Chinmayi KM, Lakshmi N, "Credit Risk Analysis in Peer-to-Peer Lending System," *IEEE International Conference on Knowledge Engineering and Applications*, 2016.
11. Z. Somayyeh, M. Abdolkarim, "Natural Customer Ranking of Banks in Terms of Credit Risk by Using Data Mining," *Jurnal UMP Social Sciences and Technology Management*, 2015.
12. A.B. Hussain, "Developing Prediction Model of Loan Risk in Banks Using Data Mining," *Machine Learning and Applications: An International Journal (MLAIJ)*, 2016.