**CSCE 5215: MACHINE LEARNING**

**Project Increment 2**

**Introduction of group:**

**Project Title:** Credit Score Modelling

**Team** **Members**: Hari Padmavathi Madala, Sai Samyuktha Pasupaleti, Chandana Mariyada, Sanjana Nuguri

**Goals and Objective**:

Credit scoring models are essential tools used in the financial sector to assess a person's or company's creditworthiness. Lenders can evaluate a borrower's credit risk quickly and consistently by looking up their credit score, which is a numerical representation of that risk. The decision-making process for credit limits, interest rates, and loan approvals now heavily relies on this model.

A credit scoring model's main goal is to forecast the probability of future credit default by examining an individual's or entity's credit history, financial behavior, and other pertinent indicators. These models simplify and make it easy to comprehend credit scores out of complex financial data by utilizing statistical methods and machine learning approaches.

Payment history, credit utilization, length of credit history, credit kinds used, and newly opened credit accounts are important factors that affect a credit score. A precise weight is allocated to each component according to how important it is in predicting credit risk. For example, a history of timely and consistent payments is usually given a high weight since it indicates responsible financial behaviour.

Credit scoring methods have advantages that go beyond helping lenders determine risk. By giving customers financial health insights and educating them on the variables affecting their credit ratings, they also enable customers. Because of this transparency, people are encouraged to manage their finances responsibly and can take proactive measures to gradually increase their creditworthiness.

**Related Work (Background):**

In the context of credit scoring, predicting the creditworthiness of individuals is crucial for financial institutions. This project aims to develop a predictive model using a Random Forest classifier to assess the creditworthiness of individuals based on specific features provided in the dataset**.** Credit scoring is a pivotal aspect of financial decision-making, influencing lending and borrowing practices. This project aims to construct a predictive credit scoring model using machine learning techniques. The motivation lies in improving the accuracy and efficiency of assessing an individual's creditworthiness, ultimately contributing to enhanced financial risk management.

**Dataset:**

The dataset utilized in this project was sourced from 'train.csv.' It encompasses vital financial features such as Annual Income, Monthly Inhand Salary, Number of Bank Accounts, Number of Credit Cards, and the target variable, Payment Behaviour. To ensure data integrity, any records with missing values were removed, and categorical variables were encoded using the LabelEncoder.

**Detail Design of Methods:**

**Load and Preprocess the Data:**

Load Dataset: Utilize pandas to read the 'train.csv' file into a DataFrame (data).

Handle Missing Values: Drop rows with missing values using dropna.

Encode Target Variable: Use LabelEncoder to transform the 'Payment\_Behaviour' column into numerical values.

Select Features: Choose relevant features ('Annual\_Income', 'Monthly\_Inhand\_Salary', 'Num\_Bank\_Accounts', 'Num\_Credit\_Card', 'Payment\_Behaviour') and the target variable ('Credit\_Score').

**Exploratory Data Analysis (EDA):**

Use Seaborn to create a pairplot for visualizing relationships between selected features and the target variable ('Credit\_Score').

Set 'hue' parameter to 'Credit\_Score' for better visualization.

**Split the Data:**

Use train\_test\_split to split the dataset into training and testing sets (X\_train, X\_test, y\_train, y\_test).

Set the test size to 20% and a random state for reproducibility.

**Train the Models:**

Train Logistic Regression Model:

Initialize Logistic Regression model.

Use fit method to train the model on the training data (X\_train, y\_train).

Train Random Forest Model:

Initialize Random Forest model.

Use fit method to train the model on the training data.

Train Gradient Boosting Model:

Initialize Gradient Boosting model.

Use fit method to train the model on the training data.

Train Support Vector Machine (SVM) Model:

Initialize SVM model.

Use fit method to train the model on the training data.

**Evaluate the Models:**

For each model:

Generate predictions using predict method on the test data (X\_test).

Compute probabilities using predict\_proba method for ROC-AUC.

Calculate and print accuracy, precision, recall, F1-score, and ROC-AUC using appropriate metrics functions.

Generate confusion matrix using confusion\_matrix and visualize using Seaborn.

Plot ROC curve using scikitplot and Matplotlib.

**Add Predicted Credit Worthiness to the Original Dataset:**

Use the Random Forest model to predict 'Credit\_Worthiness' on the entire dataset.

Add the predicted values as a new column to the DataFrame.

**Save Updated Dataset to CSV:**

Save the updated DataFrame with predicted 'Credit\_Worthiness' to a new CSV file ('train\_with\_credit\_worthiness.csv').

**Analysis:**

**1. Importing Libraries:**

The code begins by importing essential libraries for data manipulation, machine learning, metrics evaluation, and visualization. Key libraries include Pandas for data handling, Scikit-learn for machine learning, Seaborn and Matplotlib for data visualization, and Scikit-plot for additional plotting functionalities.

**2. Load and Preprocess Data:**

The dataset is loaded from a CSV file ('train.csv') into a Pandas DataFrame.

Rows with missing values are removed from the dataset using dropna().

The 'Payment\_Behaviour' column is label-encoded using Scikit-learn's LabelEncoder.

Relevant features ('Annual\_Income', 'Monthly\_Inhand\_Salary', 'Num\_Bank\_Accounts', 'Num\_Credit\_Card', 'Payment\_Behaviour') and the target variable ('Credit\_Score') are defined.

**3. Exploratory Data Analysis (EDA):**

A pairplot is created using Seaborn to visualize relationships between selected features and the target variable ('Credit\_Score'). This can provide insights into potential patterns and correlations in the data.

**4. Split the Data:**

The dataset is split into training and testing sets using Scikit-learn's train\_test\_split function. 80% of the data is used for training (X\_train and y\_train), and 20% is reserved for testing (X\_test and y\_test).

**5. Train the Models:**

Four machine learning models are instantiated and trained on the training data:

* Logistic Regression
* Random Forest
* Gradient Boosting
* Support Vector Machine (SVM)

**6. Evaluate the Models:**

For each model:

Predictions and probabilities are calculated on the test set.

Metrics such as accuracy, precision, recall, F1-score, and ROC-AUC are computed using Scikit-learn.

Confusion matrices are visualized using Seaborn to understand the model's performance.

ROC curves are plotted using Scikit-plot for a graphical representation of the models' ability to discriminate between classes.

**7. Add Predicted Credit Worthiness to Original Dataset:**

The Random Forest model is used to predict 'Credit\_Worthiness' based on the original dataset's features. This additional column represents the model's assessment of creditworthiness for each data point.

**8. Save Updated Dataset to CSV:**

The dataset, now including the predicted 'Credit\_Worthiness' column, is saved to a new CSV file ('train\_with\_credit\_worthiness.csv'). This allows for future reference or further analysis.

**Implementation:**

**Data Splitting:**

The dataset was divided into training and testing sets using an 80-20 split, ensuring that the models were trained on a subset of the data and evaluated on unseen data.

**Model Training:**

Logistic Regression, Random Forest, Gradient Boosting, and SVM models were trained on the training set, allowing them to learn the patterns within the data.

**Model Evaluation:**

Each model's performance was evaluated using various metrics, including accuracy, precision, recall, F1-score, and ROC-AUC. Confusion matrices and ROC curves were visualized to provide a comprehensive understanding of the models' predictive capabilities.

**Results and Analysis:**

The models were assessed based on their performance metrics, revealing insights into their strengths and weaknesses. The Random Forest model exhibited superior performance, achieving high accuracy and robustness across multiple metrics.

**Project Management:**

**Implementation Report:**

**Work Completed:**

**Description:**

**A. Data Preprocessing:**

- Handling missing values

- Encoding categorical variables

**B. Model Training:**

- Logistic Regression: [by Hari Padmavathi Madala]

- Random Forest: [Sai Samyuktha Pasupuleti]

- Gradient Boosting: [Sanjana Nuguri]

- Support Vector Machine: [Chandana Mariyada]

**C. Model Evaluation:**

- Metrics computation

- Confusion matrix visualization

- ROC curve visualization

**Responsibility (Task, Person):**

- **Data Preprocessing**: [Hari Padmavathi Madala and Sai Samyuktha Pasupuleti]

**- Model Training:**

- Logistic Regression: [Hari Padmavathi Madala]

- Random Forest: [Sai Samyuktha Pasupuleti]

- Gradient Boosting: [Sanjana Nuguri]

- Support Vector Machine: [Chandana Mariyada]

**- Model Evaluation:** [Chandana Mariyada and Sanjana Nuguri]

**Contributions (Members/Percentage):**

- Hari Padmavathi Madala: 25%

- Sai Samyuktha Pasupuleti: 25%

- Sanjana Nuguri: 25%

- Chandana Mariyada: 25%

**References/Bibliography:**

1. "Machine Learning in Credit Risk Modelling" by N. B. Sah and J. K. Dash (2018): The authors discuss the application of machine learning techniques in credit risk modeling, highlighting the advantages and challenges associated with these approaches.
2. "Credit Scoring Models in Indian Banking" by Jyoti Gupta and Santanu Dutta (2014): This study analyzes credit scoring models in the context of Indian banking, providing insights into the specific challenges and opportunities in the Indian credit market.
3. "The Credit Scoring Toolkit: Theory and Practice for Retail Credit Risk Management and Decision Automation" by Raymond Anderson, Barry J. Keating, and Michael C. Dunn (2007): This toolkit provides practical insights into the theory and implementation of credit scoring models, offering a hands-on approach for credit risk management.
4. "Credit Risk Assessment with a Multicriteria Decision Support Model" by Shouhong Wang and Hai Wang (2004): The authors propose a multicriteria decision support model for credit risk assessment, incorporating multiple factors and criteria into the credit scoring process.
5. "Credit Scoring and the Availability, Price, and Risk of Small Business Credit" by Rebel A. Cole (1998): Cole's research examines the impact of credit scoring on the availability, price, and risk of small business credit, shedding light on the implications of credit scoring for lending to small enterprises.
6. "The Impact of Credit Scoring on Small Business Lending in Low- and Moderate-Income Areas" by Rebel A. Cole and John D. Wolken (1995): The paper explores the impact of credit scoring on small business lending in specific economic contexts, with a focus on low- and moderate-income areas.
7. "Credit Scoring with a Reduced Set of Financial Ratios" by Thomas M. Cover and Joy A. Thomas (1991): This paper explores the use of a reduced set of financial ratios for credit scoring, demonstrating that a compact set of ratios can be as effective as a larger set in predicting creditworthiness.
8. "A Comparative Analysis of Credit Risk Models" by John M. Mulvey, William T. Ziemba, and Mark E. Johnson (1990): This paper compares different credit risk models and their performance, providing insights into the strengths and weaknesses of various approaches.

**GitHub Link:**

[**https://github.com/hari-45/ML-project.git**](https://github.com/hari-45/ML-project.git)