KIET
GROUP OF INSTITUTIONS
“Shaping Young Minds

Name of the Students:

Sujal Gupta	**(202410116100214)**
Shivam Chaudhary	**(202410116100199)**
Shubhranshu	**(202410116100205)**

Branch: MCA

Section: D

Session: 2024-2025

Submitted to: Mrs Komal Salgotra

Credit Score Prediction Using Machine Learning

Introduction:

Credit score is a fundamental metric used by financial institutions to assess an individual's creditworthiness. It serves as a numerical representation of a borrower's financial behavior and reliability in repaying debts. A higher credit score generally indicates that an individual is a responsible borrower with a lower risk of default, making them more likely to receive favorable loan terms, such as lower interest rates and higher credit limits. Conversely, a lower credit score suggests a higher risk for lenders, often leading to loan rejections, increased interest rates, or stricter borrowing conditions.

Traditionally, credit scores are determined based on historical financial data, including credit history, outstanding debts, repayment behavior, and credit utilization. These scores are typically generated using rule-based algorithms, such as FICO and VantageScore, which rely on predefined weighting systems. However, conventional methods may not always capture hidden patterns in a borrower's financial behavior, leading to potential inaccuracies in credit assessment.

With advancements in artificial intelligence and machine learning, predictive modeling has become an effective approach to enhancing credit score evaluation. Machine learning models can identify complex relationships between financial variables, uncovering insights that traditional scoring mechanisms might overlook. By analyzing key financial attributes, such as age, income, and loan amount, machine learning models can generate more precise credit score predictions, improving risk assessment for financial institutions.

In this project, we develop a machine learning model using **Linear Regression** to predict a customer's credit score based on their **age, income, and loan amount**. Linear Regression is a simple yet powerful statistical technique that establishes a relationship between independent variables (predictors) and a dependent variable (credit score). The primary objective of this model is to provide financial institutions with a data-driven approach to estimating credit scores, enabling more accurate decision-making in loan approvals, interest rate determination, and overall risk management.

By leveraging machine learning in credit scoring, this project aims to:

- Enhance the accuracy of credit score predictions.
- Reduce bias associated with traditional credit evaluation methods.
- Assist lenders in making informed financial decisions.
- Improve financial inclusion by offering alternative scoring mechanisms.

Through this project, we demonstrate how machine learning can revolutionize the credit assessment landscape, making lending processes more efficient and data-driven.

Methodology

Data Collection

The dataset contains the following features:

- **Age**: The applicant’s age in years.
- **Income**: The applicant’s annual income in currency units.
- **Loan Amount**: The total loan amount requested.
- **Credit Score** (Target variable): A numerical score representing the applicant’s creditworthiness.

The data is loaded from a CSV file using **Pandas** in Python.

Data Preprocessing

Before training the model, the data undergoes preprocessing:
✓ **Handling missing values** (if any).
✓ **Checking for outliers** and extreme values that may impact predictions.
✓ **Normalizing the data** if required to improve model efficiency.
✓ **Splitting the dataset** into training (80%) and testing (20%) sets.

Model Selection

We use **Linear Regression**, a widely used **supervised learning algorithm**, which establishes a relationship between input features and the target variable. The equation used in Linear Regression is:

y^=β0+β1x1+β2x2+β3x3+...+βnxn\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + ... + \beta_n x_ny^=β0+β1x1+β2x2+β3x3+...+βnxn

Where:

- y^\hat{y}y^ is the predicted Credit Score.
- x1,x2,x3x_1, x_2, x_3x1,x2,x3 are the input features (**Age, Income, Loan Amount**).
- β0\beta_0β0 is the intercept, and β1,β2,β3\beta_1, \beta_2, \beta_3β1,β2,β3 are coefficients that the model learns.

Model Training & Evaluation

The **Linear Regression model** is trained using the **training dataset**, and its performance is evaluated on the **testing dataset** using the following metrics:

- **Mean Absolute Error (MAE)**: Measures the average absolute difference between actual and predicted values.
- **Mean Squared Error (MSE)**: Squares the differences to penalize large errors more heavily.
- **R-squared Score (R2R^2R2)**: Measures how well the model explains the variance in Credit Score.

Data Visualization

To gain insights, we visualize the model performance using:
✓ **Actual vs. Predicted Credit Score** – A scatter plot to compare predictions with actual values.
✓ **Residual Plot** – To analyze errors and check if they follow a normal distribution.
✓ **Feature Importance** – A bar plot of the regression coefficients to show which features impact credit scores the most.

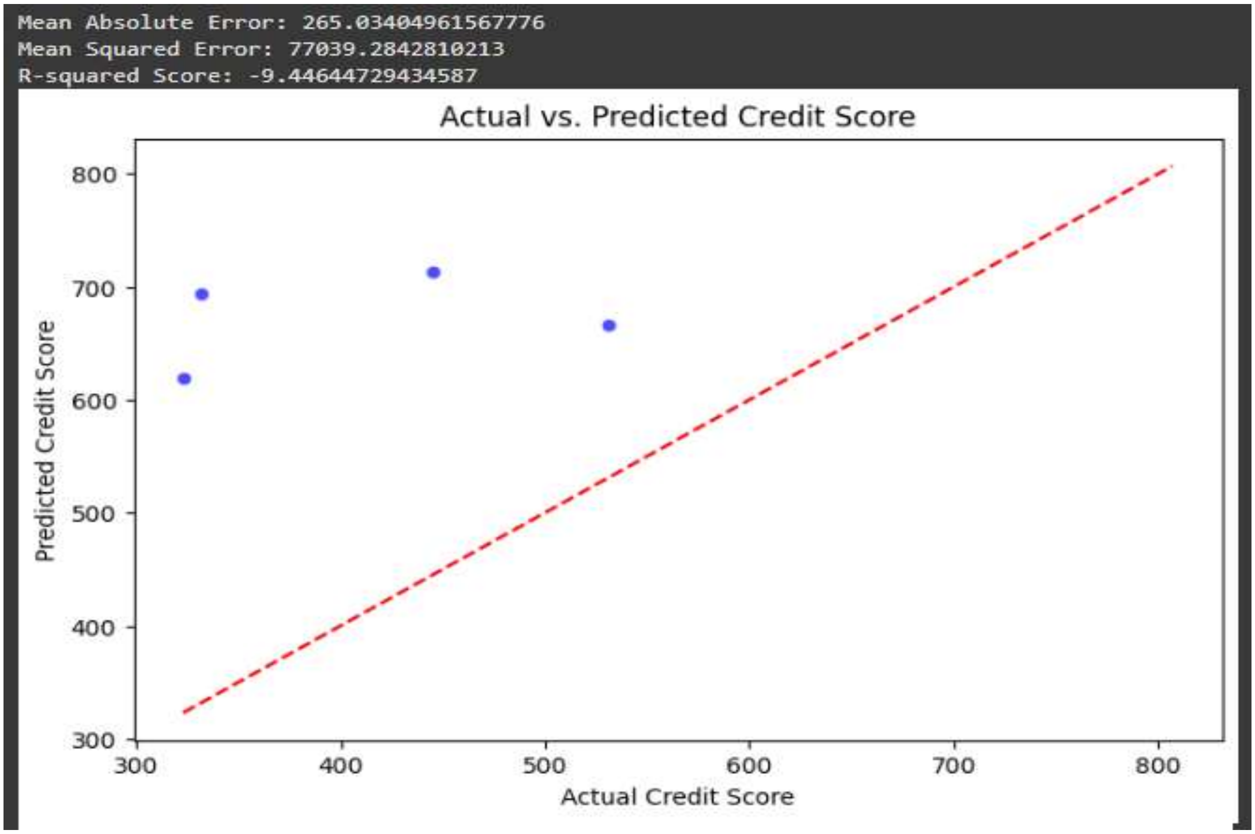
Credit Score Prediction

Once trained, the model is used to predict the credit score for new customers based on their **Age, Income, and Loan Amount**.

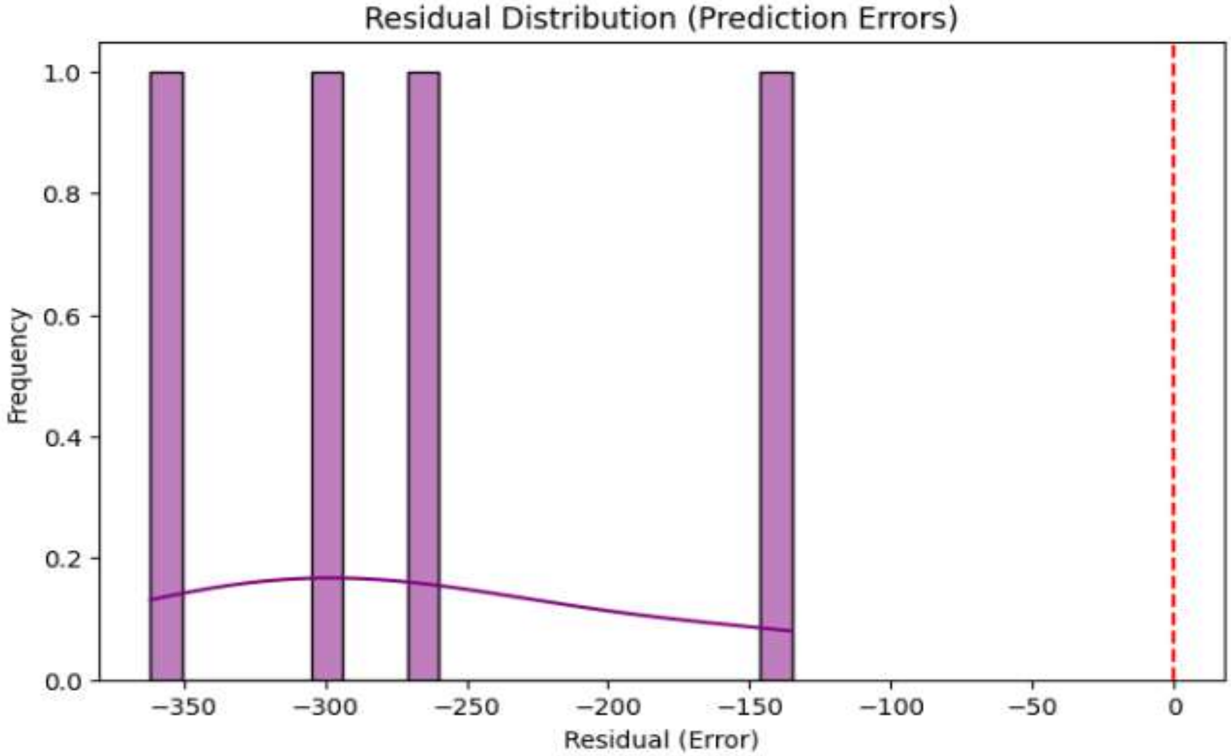
Code:

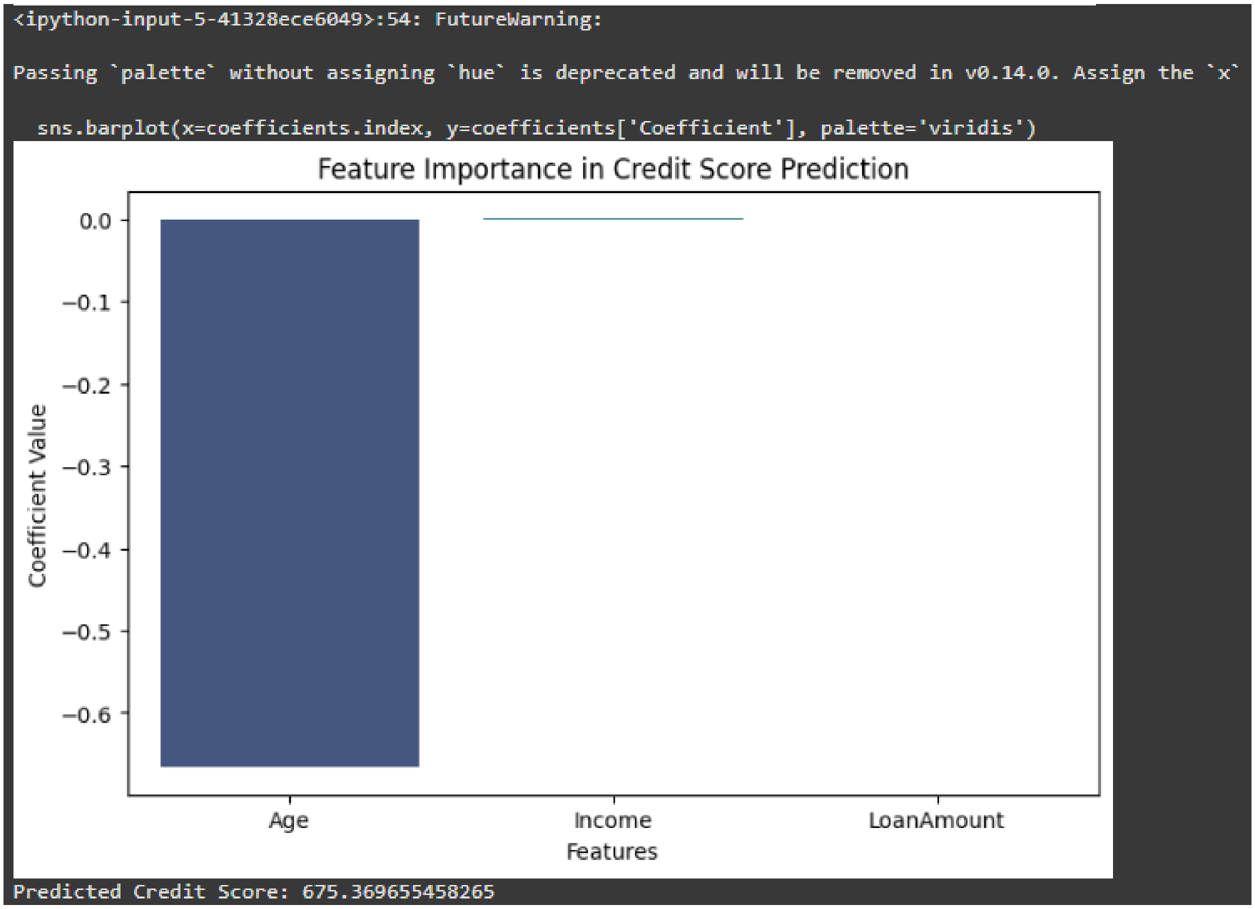
```
# Import necessary libraries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_absolute_error, mean_squared_error, r2_score

# Load the dataset
data = pd.read_csv('/content/credit_data.csv')

# Define features and target variable
X = data[['Age', 'Income', 'LoanAmount']]
y = data['CreditScore']

# Split the dataset into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Train a Linear Regression model
model = LinearRegression()
model.fit(X_train, y_train)

# Predict on test set
y_pred = model.predict(X_test)

# Evaluate the model
print("Mean Absolute Error:", mean_absolute_error(y_test, y_pred))
print("Mean Squared Error:", mean_squared_error(y_test, y_pred))
print("R-squared Score:", r2_score(y_test, y_pred))

# Visual 1: Actual vs. Predicted values
plt.figure(figsize=(8, 5))
sns.scatterplot(x=y_test, y=y_pred, color='blue', alpha=0.7)
plt.plot([y.min(), y.max()], [y.min(), y.max()], color='red', linestyle='--')  # Line of perfect predictions
plt.xlabel("Actual Credit Score")
plt.ylabel("Predicted Credit Score")
plt.title("Actual vs. Predicted Credit Score")
plt.show()

# Visual 2: Residual plot
residuals = y_test - y_pred
plt.figure(figsize=(8, 5))
sns.histplot(residuals, bins=20, kde=True, color='purple')
plt.axvline(0, color='red', linestyle='--')
plt.xlabel("Residual (Error)")
plt.ylabel("Frequency")
plt.title("Residual Distribution (Prediction Errors)")
plt.show()

# Visual 3: Feature importance (coefficients)
coefficients = pd.DataFrame(model.coef_, X.columns, columns=['Coefficient'])
plt.figure(figsize=(8, 5))
sns.barplot(x=coefficients.index, y=coefficients['Coefficient'], palette='viridis')
plt.xlabel("Features")
plt.ylabel("Coefficient Value")
plt.title("Feature Importance in Credit Score Prediction")
plt.show()

# Predict Credit Score for new data
new_data = pd.DataFrame({'Age': [30], 'Income': [70000], 'LoanAmount': [20000]})
predicted_score = model.predict(new_data)
print("Predicted Credit Score:", predicted_score[0])
```

Output:



```
Mean Absolute Error: 265.03404961567776
Mean Squared Error: 77039.2842810213
R-squared Score: -9.44644729434587
```





```
<ipython-input-5-41328ece6049>:54: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in v0.14.0. Assign the `x`

  sns.barplot(x=coefficients.index, y=coefficients['Coefficient'], palette='viridis')
```

```
Predicted Credit Score: 675.369655458265
```