# Credit Score Classification

DATE: 21/11/24

Author: Priyank Bansal
Institution: INDIAN INSTITUTE OF INFORMATION TECHNOLOGY RANCHI JHARKHAND
Supervisor/Mentor: Dr. Shivang Tripathi (Assistant Professor in CSE Department)

## I.**BRIEF SUMMARY**:

This project aims to create a machine learning model that can predict a person's credit score using various financial and personal factors. This prediction can help in making smart financial choices, like approving loans and assessing risks. The team cleaned the dataset, dealt with missing data, and coded categorical variables to get it ready for model training. They chose a Random Forest Classifier because it's tough and can handle different types of data. The dataset was split to train and test the model. The Random Forest model did a good job predicting credit score categories. Looking at feature importance showed which factors matter most for creditworthiness such as income, debt, and credit history. The project shows how machine learning can work in finance. Banks and other financial companies can use this model to make data-based decisions and lower credit risk. In the future, the team might try advanced methods like deep learning and add more relevant features to make the model even better at predicting.

## II.**INTRODUCTION**:

### 1.Background

Credit scores play a key role in showing a person's financial health and their ability to get various financial products. When banks and lenders can predict credit scores , it helps both them and their customers. By knowing what affects creditworthiness, these companies can make smart choices about giving loans and figuring out risks. For people looking to borrow, having a good idea of their credit score lets them plan their finances better and work on boosting their credit rating.

### 2.Objectives

Data Preparation: To clean and prepare the dataset so it's accurate and consistent.

Model Selection and Training: To pick the right machine learning method (Random Forest Classifier) and teach it using the prepared data.

Model Evaluation: To check how well the model works using key measures like accuracy, precision, recall, and F1-score.

Feature Importance Analysis: To find out which factors have the biggest effect on credit scores.

### 3.Scope

This project aims to build a machine learning model that predicts credit scores from a given dataset. It covers:

1.Looking at and cleaning up the data
2.Creating and choosing useful features
3.Training and fine-tuning the model
4.Testing and understanding the model

The project doesn't cover putting the model into real-world use or dealing with predictions in real-time.

### 4.Problem Statement

The primary challenge is to develop a robust machine learning model that can accurately predict an individual's credit score category based on a set of financial and personal attributes. The model should be able to generalize well to unseen data and provide meaningful insights into the factors driving creditworthiness.

## III.**LITERATURE REVIEW / BACKGROUND STUDY:**

Credit Scoring: A Quick Look:

Credit scoring helps banks figure out if someone can pay back a loan. It looks at things like how much money they make, what they owe, and if they've paid bills on time before. This helps predict if they might not pay back a loan. In the past, banks used math methods like logistic regression to do this. But now, with computers getting smarter, they're using new ways to score credit.

Machine Learning in Credit Scoring:

Machine learning algorithms have caught the eye of many in credit scoring. This comes from their skill in dealing with big and tricky data sets. Research has looked into how different machine learning methods can help such as:

Decision Trees: These create a map of choices and what might happen. People can understand them, but they might fit the data too.

Random Forest: This method puts together many decision trees to get better results and avoid fitting the data too. It's a go-to choice for credit scoring because it's tough and can work with different kinds of data.

Support Vector Machines (SVM): SVM is a strong tool for sorting data. It finds the best line to split data points. It works well when there's lots of data to consider.

Neural Networks: These models, inspired by the human brain, can learn complex patterns from large datasets. Deep learning techniques, such as convolutional neural networks (CNN) and recurrent neural networks (RNN), have shown promising results in credit scoring.

XGBoost: A gradient boosting framework that iteratively improves the model by adding new weak learners. It is known for its efficiency and accuracy.

Challenges and Future Directions:

While machine learning has the potential to revolutionize credit scoring, several challenges remain:

1.Data Quality and Quantity: High-quality and sufficient data is essential for training accurate models. Data imbalance and missing values can impact performance.

2.Model Interpretability: Complex models like deep neural networks can be difficult to interpret, hindering trust and transparency.

3.Ethical Considerations: Models should be fair and unbiased, avoiding discriminatory practices.

4.Evolving Financial Landscape: The financial landscape is constantly changing, and models need to adapt to new trends and regulations.

Future research directions include exploring advanced techniques like explainable AI to improve model interpretability, incorporating alternative data sources (e.g., social media, mobile data), and developing hybrid models that combine traditional statistical methods with machine learning

## IV. **Methodology:**

Approaching:

A structured approach is used to address the credit score classification issue:

1.Data acquisition and pre-processing:
Compile a data set that contains relevant financial and personal characteristics of an individual.
Data was cleaned to control missing values, outliers, and inconsistencies.
Perform feature engineering to create new features or modify existing features for better model performance.

2.Model selection and training:
Random forest classification was chosen as the main algorithm because of its robustness in handling mixed data types.
Divide the dataset into training and testing sets.
Train the model on the training set using appropriate hyperparameters.

3.Model evaluation:
Evaluate the model's performance on the test set using metrics such as precision, precision, recall, and F1 score.
Analyze the model's predictions to identify strengths and weaknesses.

4.Feature importance analysis:
Use the built-in feature importance feature of the random forest model to determine the most influential factors in predicting credit scores way:
1.Data pre-processing:
Imputation is not available (e.g. mean, median, mode, or advanced techniques such as KNN imputation).
Identification and handling of abnormalities (e.g., capping, flooring, or removal).
Scaling features (e.g. normalization or standardization).
Coding of categorical variables (e.g., one-hot coding label encoding).
2.Ideal training:
Hyperparameter tuning (e.g. table lookup random search).
Model training and validation.
3.Model evaluation: Confusion
Matrix.
Classification report.
ROC graph and AUC-ROC score.
4.Feature importance analysis:
The nature of random forests is an important feature.

Data Collection:
The dataset used in this project was obtained from STATSO(Kaggle) Contains a wide range of financial dataset for case studies.

Project planning:
1. Data Acquisition and Exploration:
Collect and clean datasets.
Explore the data to understand features and identify potential issues.
2.Data preprocessing and feature engineering:

Data preprocessing to handle missing values and outliers.
Create new properties or edit existing properties to improve model performance.
3.Model selection and training:
Choose the right machine learning algorithm. (random forest).
Divide the dataset into training and testing sets.
Train the model in the training set.
4.Model evaluation and tuning:
Model performance was evaluated on the test set.
Fine-tune model hyper-parameters to improve accuracy.
5.Feature importance analysis:
Identify the most influential factors in predicting credit scores. Implementation/Design:

## V. **System architecture**:

Basic machine learning pipeline Including the provision of information Pre-processing, model training, evaluation and deployment...
The system architecture for this credit score classification project follows a standard machine learning pipeline:

1.Data Ingestion:
Load the dataset in the appropriate format (e.g. CSV, JSON).
2.Data pre-processing:
Cleaned the data to deal with missing values and outliers.
Perform feature engineering to create new features or modify existing features.
3.Model training:
Divide the dataset into training and testing sets.
Train a random forest classifier on the training set.
4.Model evaluation:
Evaluate the model's performance on the test set using various metrics.
5. Model Deployment:
Deploy the trained model to a production environment (optional).

Operation details:
1.Data pre-processing:
Python libraries such as Pandas and NumPy are used for data manipulation and analysis.
Missing values were corrected using imputation techniques.
Numerical properties are scaled to a normal range.
Category features encoded using single-hot encoding or label encoding.
2.Model training:
The Scikit-learn library is used to implement the Random Forest Classifier. Experiment with different extreme parameters (e.g. number of trees, maximum depth) to optimize performance.
3.Model evaluation:
Accuracy, precision, recall, and F1 score were calculated to evaluate the model performance.

Model performance was visualized using confusion matrices and ROC curves.
4.Feature importance analysis: The feature_importances_ feature of the Random Forest model is used to identify the most influential features.

Technology Used:
Python: A programming language for data analysis and model development.
Pandas and NumPy: Data management and analysis libraries
Matplotlib/Seaborn: Data visualizing trends, relationships, and patterns libraries.
Scikit-learn: A machine learning library for model training and evaluation.
Matplotlib and Seaborn: Data visualization libraries.
Jupyter Notebook: An interactive computing environment for data exploration and analysis.
Google Colab: An online, cloud-based platform that allows you to data analysis ,data exploration and visualization of machine learning models.

Challenges and solutions:
1.Unbalanced datasets: Addresses the issue of unbalanced classes using techniques such as oversampling. Sampling is too little or class weighting
2.Feature Engineering: Using different combinations of features. and adjustments to improve model performance.
3.Hyperparameter tuning: Use network search or random search to find the optimal hyperparameters of the random forest model.
4.Model Interpretability: Techniques such as SHAP values are used to describe model predictions and identify the most influential features.

VI. **<u>Results and analysis result</u>**: Model performance:

A random forest classifier is obtained. .8063 on the test set. This shows that the model predicts 80.63% of credit score correctly

Feature Importance:

The following characteristics were identified as most influential in determining credit worthiness. Annual

Income: Higher income often indicates a better ability to repay debt.

Debt-to-income ratio: A lower debt-to-income ratio indicates a lower financial burden.
Length of Credit History: Generally, a longer credit history with a good payment history is beneficial.
Number of credit checks: Frequent credit checks may indicate a higher risk of default.
Credit Utilization Ratio: A low credit utilization ratio indicates responsible use of credit.
[Insert bar chart or table showing feature importance scores][continuously improvent in the report] Table

1: Classification report

```
                    Precision  recall  f1-score   support

Good                 0.77      0.77      0.77      1827
Poor                 0.79      0.83      0.81      3023 Standard
0.82      0.80      0.81       5397
 accuracy                             0.80
10247 macro avg        0.80      0.80      0.80
10247
weighted avg    0.80      0.80      0.80     10247
```

Analysis:
The random forest model shows strong performance in predicting debt scores. The importance of the identified characteristics is consistent with financial intuition. It emphasizes the model's ability to capture relevant patterns in the data.

However, it's important to note that model performance can be affected by factors such as data quality. Feature Engineering hyperparameter tuning, etc. Continuous monitoring and improvement of models is essential to maintaining accuracy and reliability in a dynamic financial environment.
Further analysis and experimentation may involve exploring more advanced techniques such as deep learning. Integration of additional features or using external data sources to increase the model's predictive ability...

VII. **Discussion:** Interpretation:

The results show that the random forest classifier is a suitable model for predicting credit scores. The importance of the identified characteristics is consistent with financial intuition. This indicates that the model has detected meaningful patterns from the data. Higher annual income Decreasing debt-to-income ratio Longer credit history Fewer credit checks and the credit utilization ratio decreased They are all strong indicators of creditworthiness.
Comparison:
The research results are consistent with previous credit scoring research, which found that these factors are important determinants of creditworthiness. The performance of the random forest model is comparable to other machine learning models. used in credit scoring which shows effectiveness in this area...

Limitations:
The following limitations were encountered during the project.

Data Quality: Data quality can have a significant impact on model performance. Inaccurate or missing data can lead to biased results.
Feature Engineering: The effectiveness of feature engineering techniques may vary depending on the specific data set and problem.
Ideal Complexity: Although random forest is a powerful algorithm, But it can be computationally expensive for large data sets.
Interpretability: Although feature importance analysis provides insights, But interpreting complex models such as deep neural networks can be challenging.

VIII. **Conclusion:** summarize:

The aim of this project is to develop a machine learning model to predict credit scores based on various financial and personal characteristics. A random forest classifier was used to train the model. and achieved a satisfactory level of accuracy in predicting credit score classes. Analysis of the importance of characteristics shows that annual income Credit-to-income ratio Length of credit history Number of credit checks and use of credit Ratio...the most influential factor in determining credit worthiness...

Recommendations:

Explore advanced techniques: Consider using advanced techniques such as deep learning or clustering methods. to further improve the model's performance. Include external data : Include additional data sources such as social media or mobile data. To increase the feature area

Focus on interpretability: Use techniques like SHAP or LIME to improve the interpretability and interpretability of the model.

Continuous monitoring and retraining: Regularly monitor model performance and retrain as needed to adapt to changing trends and patterns.

Concluding remarks:

The successful development of this credit score prediction model demonstrates the potential of machine learning in the financial sector. Accurate creditworthiness assessments enable financial institutions to make informed decisions, reduce risk, and improve customer satisfaction. However, it is important to address the limitations and ethical considerations associated with learning models. of the machine To ensure fairness and transparency of credit scoring practices.

**|| Thanks to Reading ||**
**|| The END ||**