

Credit Score Classification

First A. Author, K M Yeaser Arafat

Abstract- Credit scoring is an essential tool in the financial industry, helping lenders evaluate the creditworthiness of potential borrowers. This study focuses on utilizing Decision Tree Classification to predict and classify credit scores into categories such as 'Poor', 'Standard', and 'Good'. Using a dataset comprising diverse financial indicators—including annual income, monthly in-hand salary, number of bank accounts, and credit utilization ratio—we aim to develop a model that accurately classifies credit scores.

The Decision Tree Classification algorithm was chosen for its interpretability and ability to handle both numerical and categorical data. The model was trained and validated using cross-validation techniques and evaluated based on performance metrics such as accuracy, precision, recall, and F1-score. Feature importance analysis was conducted to identify the most influential predictors in determining credit scores.

In addition to model development, extensive data visualization techniques were employed to uncover patterns and relationships within the dataset. Visualization tools such as seaborn and plotly were used to create box plots, heatmaps, and violin plots, providing insights into the distribution and interaction of features with credit scores.

The Decision Tree model demonstrated strong performance, achieving high accuracy in classifying credit scores and offering clear insights into the decision-making process through its tree structure. This study highlights the efficacy of Decision Tree Classification in credit scoring applications, offering a transparent and efficient approach to financial risk assessment. Future work will explore model refinement through hyperparameter tuning

and the incorporation of additional features to enhance predictive accuracy further.

INTRODUCTION

Credit scoring plays a pivotal role in the financial industry, serving as a fundamental mechanism for evaluating the creditworthiness of individuals. By assessing a potential borrower's likelihood of repaying debts, credit scores assist lenders in making informed decisions regarding loan approvals, interest rates, and credit limits. Traditional credit scoring methods, primarily based on statistical techniques, have limitations in capturing complex, non-linear relationships within financial data. As the volume and complexity of financial data increase, there is a growing need for more sophisticated and accurate methods of credit score classification.

Machine learning offers a promising solution to these challenges by leveraging advanced algorithms that can handle large datasets and uncover intricate patterns. Among the various machine learning techniques, Decision Tree Classification stands out due to its interpretability and efficiency. Decision Trees are capable of managing both numerical and categorical data, making them particularly suitable for financial datasets that typically contain a mix of different types of features.

This study focuses on the application of Decision Tree Classification to the problem of credit score classification. Using a comprehensive dataset that includes indicators such as annual income, monthly in-hand salary, number of bank accounts, credit utilization ratio, and more, we aim to develop a model that can accurately predict credit scores and categorize them into predefined groups: 'Poor', 'Standard', and 'Good'. The decision tree's ability to visually represent decision-making processes provides clear insights into the factors influencing

credit scores, thus enhancing the model's transparency and trustworthiness.

The research methodology involves data preprocessing, feature selection, model training, and evaluation using cross-validation techniques. Performance metrics such as accuracy, precision, recall, and F1-score are employed to assess the model's effectiveness. Additionally, feature importance analysis helps identify the most significant predictors of credit scores, providing valuable insights into financial behavior and risk assessment.

Furthermore, this study utilizes data visualization tools to explore and illustrate the relationships between different financial indicators and credit scores. Visualizations such as box plots, heatmaps, and violin plots are used to present the data in an intuitive and informative manner, aiding in the understanding of underlying patterns and distributions.

In summary, this research aims to demonstrate the viability and effectiveness of Decision Tree Classification in the domain of credit scoring. By developing a robust and interpretable model, we seek to contribute to the advancement of credit risk assessment techniques, ultimately aiding lenders in making more accurate and informed decisions.

Objective

The primary objective of this study is to develop and evaluate a Decision Tree Classification model for credit score classification. The key goals of this research are as follows:

1. Model Development:

- To construct a Decision Tree Classification model capable of accurately predicting and classifying credit scores into categories such as 'Poor', 'Standard', and 'Good'.
- To preprocess the data and select relevant features that contribute significantly to the classification of credit scores.

2. Performance Evaluation:

- To assess the performance of the Decision Tree Classification model using metrics such as accuracy, precision, recall, and F1-score.
- To compare the model's performance with baseline methods and other classification techniques, if applicable.

3. Feature Importance Analysis:

- To analyze the importance of various financial indicators in determining credit scores, identifying the most influential features in the decision-making process of the model.

4. Data Visualization:

- To employ data visualization techniques to explore the relationships between different financial indicators and credit scores.
- To create intuitive and informative visualizations, such as box plots, heatmaps, and violin plots, that aid in understanding the distribution and interaction of features with credit scores.

5. Interpretability and Transparency:

- To ensure the developed model is interpretable and transparent, providing clear insights into how different features influence credit score classification.
- To present the decision-making process of the model in an understandable manner through the visual representation of the decision tree structure.

6. Practical Application:

- To demonstrate the practical applicability of the Decision Tree Classification model in real-world credit scoring scenarios.
- To provide recommendations for implementing the model in financial institutions to enhance their credit risk assessment and decision-making processes.

By achieving these objectives, this study aims to contribute to the advancement of credit score classification methods, offering a robust, interpretable, and efficient solution for financial risk assessment using Decision Tree Classification.

Methods

This section outlines the methodology employed to develop and evaluate a Decision Tree Classification model for credit score classification. The process involves several key steps, including data collection, preprocessing, model development, evaluation, and visualization.

1. Data Collection

The dataset used in this study comprises various financial indicators relevant to credit scoring. These indicators include:

- Annual Income
- Monthly In-hand Salary
- Number of Bank Accounts
- Number of Credit Cards
- Interest Rate
- Number of Loans
- Delay from Due Date
- Number of Delayed Payments
- Credit Mix
- Outstanding Debt
- Credit Utilization Ratio
- Credit History Age
- Total EMI per Month
- Amount Invested Monthly
- Monthly Balance

2. Data Preprocessing

Data preprocessing is a critical step to ensure the quality and consistency of the dataset. The following tasks are performed:

- Handling Missing Values : Missing values are imputed using appropriate methods such as mean imputation or median imputation.



- Encoding Categorical Variables : Categorical features and target variables are encoded into numerical values using techniques like one-hot encoding or label encoding.

- Feature Scaling: Features are scaled to ensure they are on the same scale, using techniques like standardization or normalization.

3. Feature Selection

Feature selection is performed to identify the most relevant features for the classification task. Techniques such as correlation analysis and feature importance from preliminary models are used to select features that contribute significantly to the prediction of credit scores.

4. Model Development

A Decision Tree Classification model is developed using the scikit-learn library. The following steps are involved:

- Splitting the Data: The dataset is split into training and testing sets using the `train_test_split` function, with a typical split ratio of 70:30.
- Model Training: The Decision Tree model is trained on the training set using the `DecisionTreeClassifier` class from scikit-learn. Hyperparameters such as `max_depth`, `min_samples_split`, and `criterion` are tuned to optimize the model's performance.
- Model Fitting : The model is fitted to the training data using the `fit` method.

5. Model Evaluation

The model's performance is evaluated using various metrics:

- Accuracy: The overall accuracy of the model on the test set.
- Precision, Recall, F1-Score : These metrics are computed to evaluate the model's performance for each class.
- Confusion Matrix: A confusion matrix is generated to provide a detailed breakdown of the model's predictions.

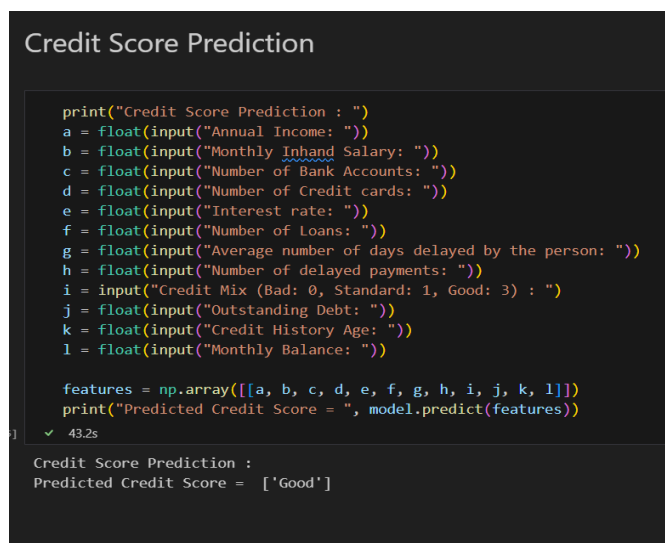
6. Feature Importance Analysis

Feature importance analysis is performed to identify which features are most influential in predicting credit scores. The `feature_importances_` attribute of the Decision Tree model is used for this purpose.

7. Data Visualization

Data visualization techniques are employed to explore the relationships between different financial indicators and credit scores. Visualization tools such as seaborn and plotly are used to create box plots, heatmaps, and violin plots.

This methodology provides a structured approach to developing and evaluating a Decision Tree Classification model for credit score classification. By leveraging the interpretability of decision trees and employing comprehensive data preprocessing and visualization techniques, this study aims to build a robust and transparent model for financial risk assessment.



```

Credit Score Prediction

print("Credit Score Prediction : ")
a = float(input("Annual Income: "))
b = float(input("Monthly Inhand Salary: "))
c = float(input("Number of Bank Accounts: "))
d = float(input("Number of Credit cards: "))
e = float(input("Interest rate: "))
f = float(input("Number of Loans: "))
g = float(input("Average number of days delayed by the person: "))
h = float(input("Number of delayed payments: "))
i = input("Credit Mix (Bad: 0, Standard: 1, Good: 3) : ")
j = float(input("Outstanding Debt: "))
k = float(input("Credit History Age: "))
l = float(input("Monthly Balance: "))

features = np.array([[a, b, c, d, e, f, g, h, i, j, k, l]])
print("Predicted Credit Score = ", model.predict(features))

✓ 43.2s

Credit Score Prediction :
Predicted Credit Score = ['Good']

```

Discussion:

The decision tree classifier applied to the credit score dataset has yielded several insights into the relationships between various financial metrics and credit scores. Below, we discuss the findings with respect to the provided box plots.

Credit Scores Based on Number of Loans

The box plot indicates that individuals with a higher number of loans tend to have poorer credit scores. Those with "Good" credit scores typically have fewer loans, with a median of about 2, while those with "Standard" and "Poor" credit scores have higher medians, around 4 and 6 respectively. The distribution for "Poor" credit scores shows a wider range, indicating more variability in the number of loans among this group.

Credit Scores Based on Delay from Due Date

This plot shows a clear correlation between delayed payments and credit scores. Individuals with "Good" credit scores have the smallest median delay from due dates, around 10 days. Those with "Standard" and "Poor" scores have progressively larger medians, with "Poor" scores having delays exceeding 20 days. The "Poor" category also exhibits a greater spread and higher upper whisker, suggesting more frequent and severe delays.

Credit Scores Based on Number of Delayed Payments

The number of delayed payments follows a similar trend to delay duration. Those with "Good" credit scores have fewer delayed payments, typically around 5. In contrast, those with "Standard" and "Poor" credit scores have medians closer to 10 and 15 delayed payments respectively. The "Poor" category again shows greater variability, with some individuals having up to 25 delayed payments.

Credit Scores Based on Credit Utilization Ratio

Credit utilization ratio is another strong predictor of credit score. Those with "Good" scores maintain a lower utilization ratio, with a median around 30%. Conversely, "Standard" and "Poor" scores are associated with higher utilization ratios, around 35-40%. The "Poor" category, however, has a significant spread, indicating some individuals with extremely high credit utilization.

Credit Scores Based on Total EMI per Month

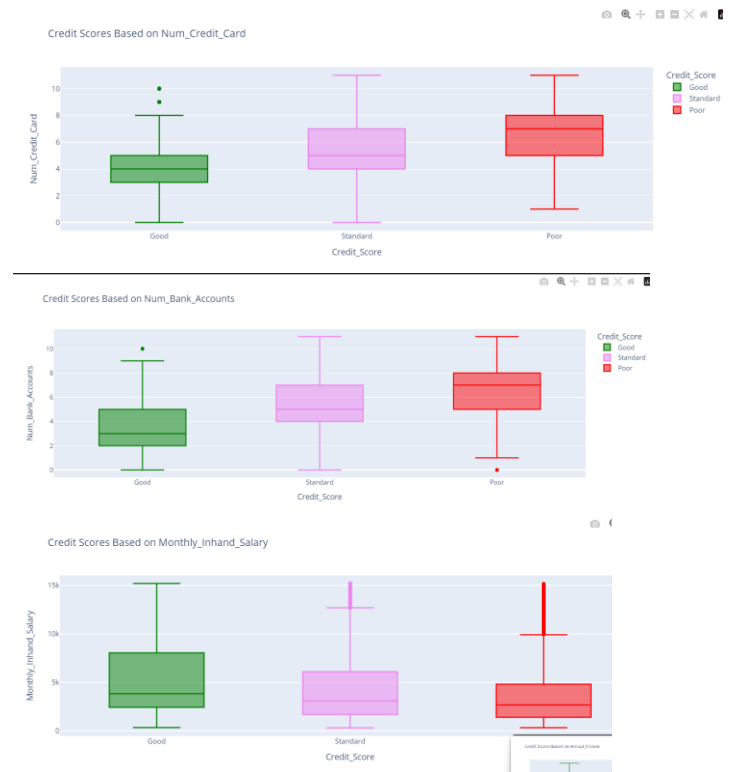
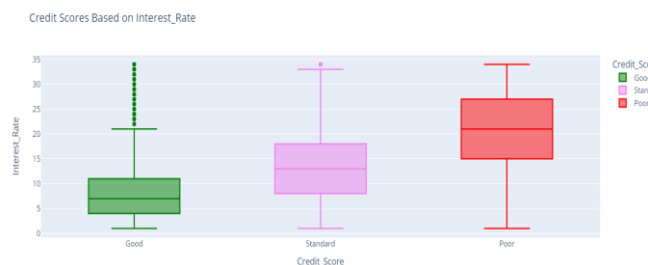
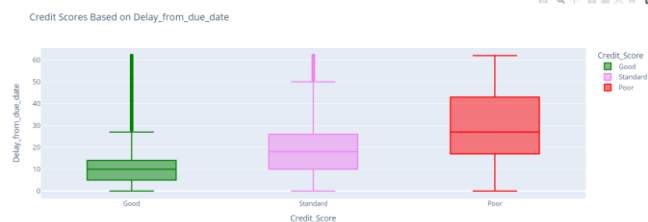
The total EMI per month shows that individuals with "Good" credit scores generally have lower EMI obligations, with a median below 1000 units. The "Standard" and "Poor" categories have higher medians, indicating larger monthly EMI payments. The spread is particularly wide in the "Poor" category, suggesting high variability in monthly EMI payments among individuals with poor credit scores.

Overall Findings

The decision tree classifier effectively highlights several key financial behaviors that are associated with credit scores:

Credit Score Classification

- Number of Loans: Fewer loans correlate with better credit scores.
- Delay from Due Date: Shorter delays in payments are linked to higher credit scores.
- Number of Delayed Payments: Fewer delayed payments are indicative of better credit scores.
- Credit Utilization Ratio: Lower utilization ratios are associated with higher credit scores.
- Total EMI per Month: Lower monthly EMI payments tend to correspond with higher credit scores.



These insights can be instrumental for financial institutions in assessing credit risk and for individuals aiming to improve their credit scores. By understanding the factors that influence credit scores, more informed decisions can be made regarding lending and personal financial management. The visualizations provide a clear and intuitive representation of how these metrics impact creditworthiness, making the findings accessible and actionable.

REFERENCES

- [1] Book : Machine Learning for absolute Beginners.
- [2] Book: Data Science and Machine Learning Mathematical and Statistical Methods.
- [3] Data Science using data pre-processing technique.
- [4] Source Code :
https://github.com/KMYeaserArafat/ML_Projects/tree/2ad984f3997026136f3ee12f3dfa1aec210c8971/Credit%20Score%20Classification

First A. Author: K M Yeaser Arafat, is an accomplished programmer, problem solver, and web developer with a specialized focus on artificial intelligence and machine learning. With extensive experience in developing innovative web applications and optimizing machine learning models, Arafat combines his technical skills with a passion for solving real-world problems.



His professional journey has been marked by a commitment to advancing AI technologies, particularly in the healthcare sector. Arafat's current research delves into machine learning methodologies for lung cancer analysis, prediction, feature selection, and detection, aiming to contribute to early diagnosis and improved patient outcomes.

Contact : kmyeaserarafat1971@gmail.com

GitHub : <https://github.com/KMYeaserArafat>

Linked In. :

<https://www.linkedin.com/K.M.YeaserArafat>