**Enhancing Credit Score Prediction:**   
Leveraging Random Forests and Data Visualization for Financial Insights



**Himansu Agrawal**

Mail – [himansubansal1701@gmail.com](mailto:himansubansal1701@gmail.com)

GitHub - <https://github.com/HIMANSUAGRAWAL>

LinkedIn – <https://www.linkedin.com/in/himansu-agrawal-45410333b/>

X - <https://x.com/HimansuBan73216>

Instagram - <https://www.instagram.com/_himansubansal_/>

Content

Abstract

1. Introduction
2. Overview of the project
3. The primary contribution of the work
4. Motivation of the Research
5. Related Work
6. Proposed Methodologies
7. Dataset Description
8. Data Preprocessing
9. Methodology
10. Evaluation of Matrices
11. Result and Discussion
12. Conclusion
13. References

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Abstract**   
This research focuses on the classification of credit scores using machine learning techniques. Credit score prediction is a crucial task in the financial industry as it helps in assessing the creditworthiness of individuals. The study employs a Random Forest Classifier to categorize individuals' credit scores into three classes: Poor, Standard, and Good. The model is trained on a dataset that includes various financial attributes such as annual income, monthly in-hand salary, number of bank accounts, and others. The dataset is pre-processed to handle missing values and categorical features are encoded to numerical values. The model's performance is evaluated using accuracy, classification reports, and confusion matrix metrics. The results demonstrate that the proposed model effectively classifies credit scores, providing valuable insights into factors influencing credit ratings. The study highlights the importance of accurate credit score prediction in reducing financial risks for lending institutions and contributes to the development of robust credit scoring models.

1. **Introduction**
2. *Overview of the Model*

The study aims to build a machine learning model for predicting credit scores, which are critical for determining an individual's creditworthiness. The credit score classification model leverages a Random Forest Classifier, a robust ensemble learning method that combines the predictions of multiple decision trees to improve accuracy and prevent overfitting.

1. *The Primary Contribution of the Work*

The primary contribution of this research lies in developing a comprehensive machine-learning model that effectively categorizes credit scores based on a variety of financial and behavioural features. By evaluating multiple aspects such as income, debt, and credit history, the model offers insights into the key factors influencing credit scores. The study also provides a detailed analysis of the dataset, preprocessing steps, and evaluation metrics used to validate the model's performance.

1. **Motivation of the Research**

The motivation behind this research stems from the need to automate and enhance the accuracy of credit score predictions. Financial institutions rely heavily on credit scores to make lending decisions, and inaccuracies can lead to significant financial risks. By utilizing machine learning techniques, this research aims to improve the precision of credit score classifications, thereby assisting lenders in making more informed decisions.

1. **Related Work**

Previous research in credit scoring has primarily focused on traditional statistical methods such as logistic regression and linear discriminant analysis. While these methods are effective, they often fall short in handling complex, non-linear relationships between variables. Recent advancements in machine learning have led to the development of more sophisticated models, such as Support Vector Machines (SVM), Decision Trees, and ensemble methods like Random Forests. This study builds on these advancements by employing a Random Forest Classifier, which has shown promise in various classification tasks due to its ability to handle large datasets and reduce overfitting.

1. **Proposed Mythologies**
2. *Dataset Description*

The dataset used in this study comprises various features related to an individual's financial behaviour and status, including annual income, monthly in-hand salary, number of bank accounts, number of credit cards, interest rates, number of loans, and more. The target variable, 'Credit Score,' is categorized into three classes: Poor, Standard, and Good.

1. *Data Preprocessing*

Data preprocessing is a crucial step to ensure the quality and consistency of the dataset. The steps involved include:

* Handling Missing Values:

Missing values are dropped or imputed depending on the context.

* Encoding Categorical Variables:

Non-numeric columns are converted into numeric form using Label Encoding for features with limited categories and OneHotEncoding for those with multiple categories.

* Feature Scaling:

This step is crucial for models sensitive to the scale of input data, although not explicitly mentioned here, it might be considered depending on the methodology used.

1. *Methodology*

The proposed model utilizes a Random Forest Classifier. The feature set is composed of financial attributes like annual income, credit utilization ratio, and others. The target variable is the 'Credit Score,' which is encoded into numeric labels. The model is trained using the train-test split method, and the performance is evaluated on the test set.

1. *Evaluation Matrix*

The performance of the model is assessed using the following metrics:

* Accuracy Score:

Measures the proportion of correct predictions.

* Classification Report:

Provides detailed metrics like precision, recall, and F1-score for each class.

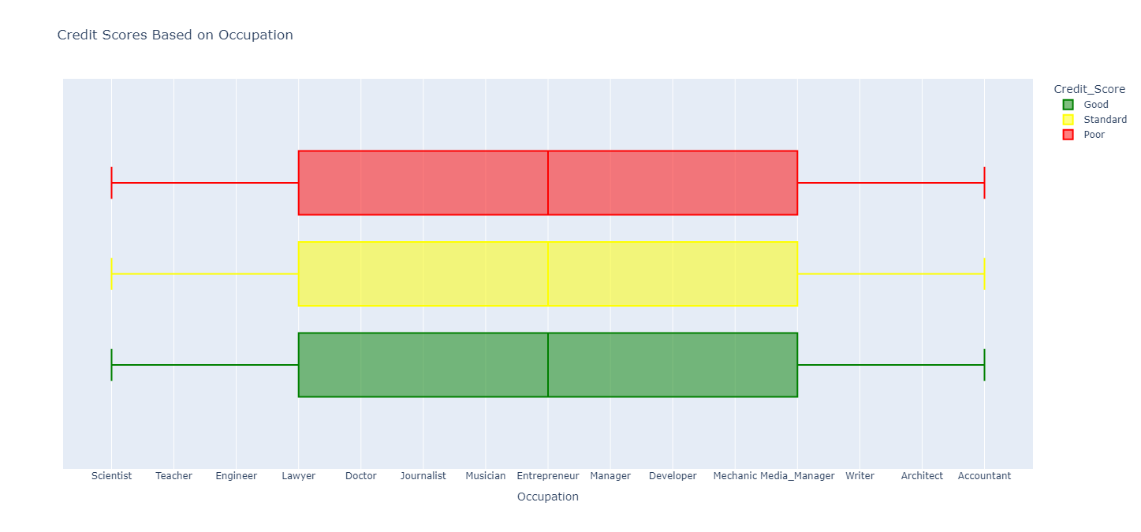
* Confusion Matrix:

Visualizes the model's performance by showing the correct and incorrect predictions for each class.

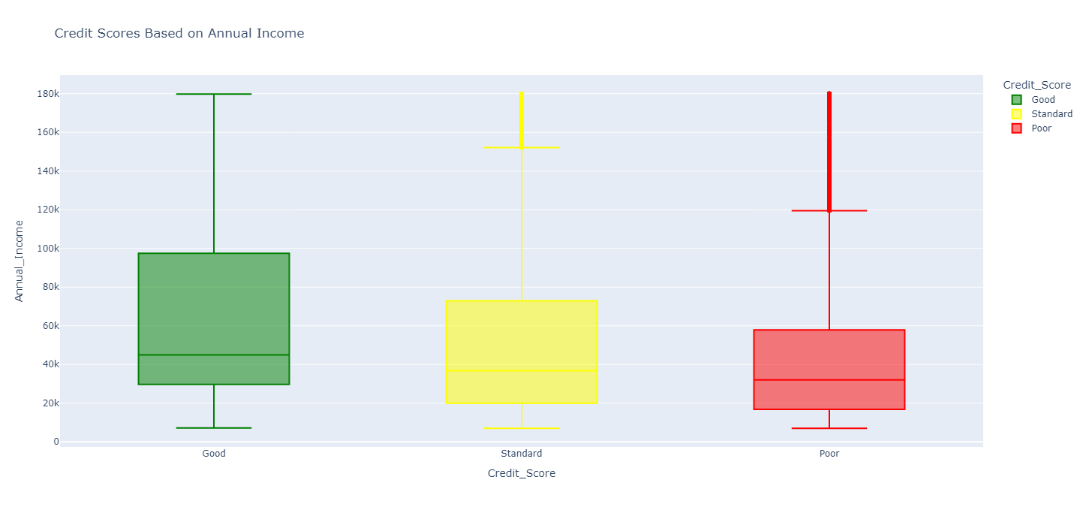
1. **Result and Discussion**
2. Visualization of Data and Feature Impact

* Credit Scores Based on Occupation:

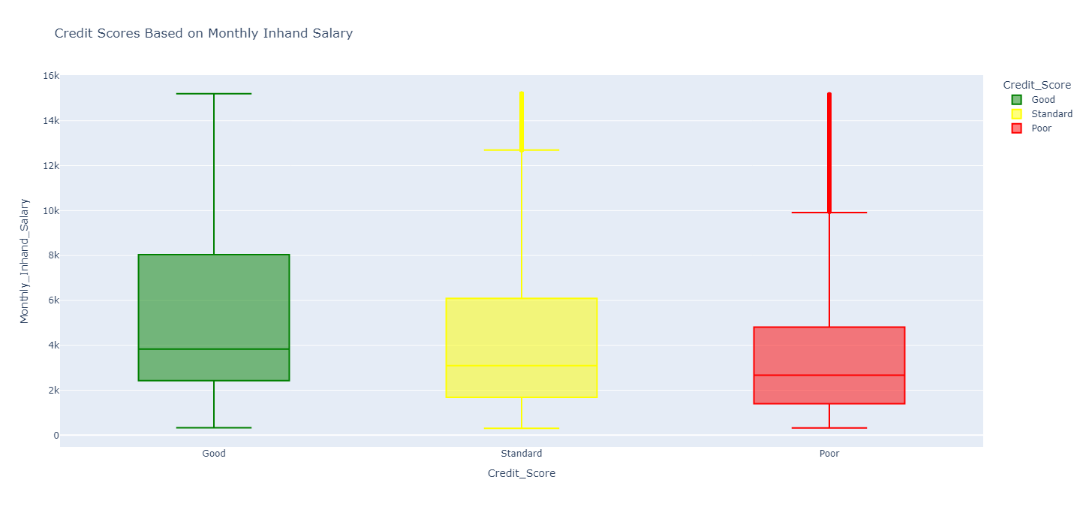
Box Plot: Visualizes the distribution of credit scores across different occupations, highlighting variations and outliers.



* Credit Scores Based on Annual Income:

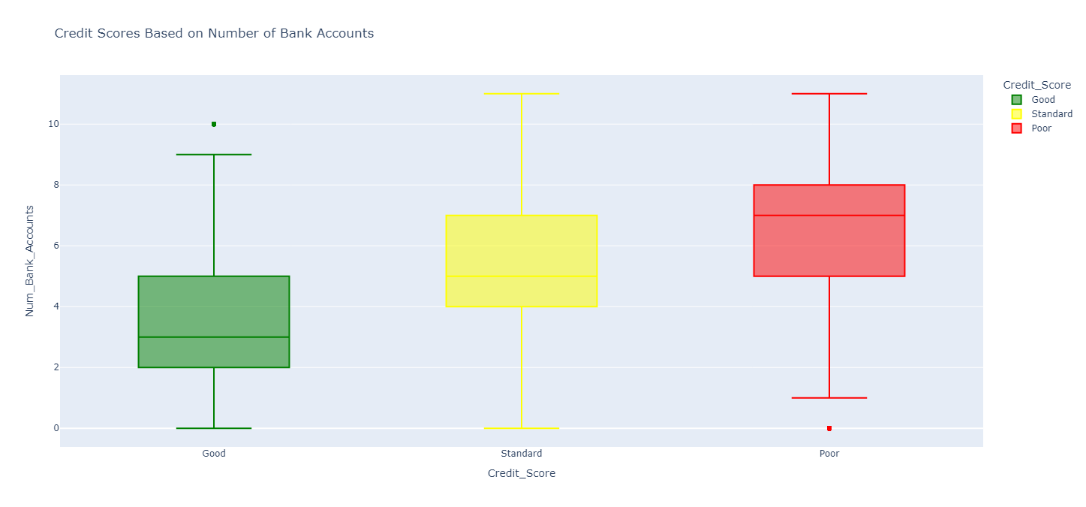
Box Plot: Shows how credit scores vary with annual income, indicating potential correlations between income levels and credit scores.

* Credit Scores Based on Monthly Inhand Salary:

Box Plot: Examines the impact of monthly salary on credit scores, providing insights into how regular earnings influence creditworthiness.

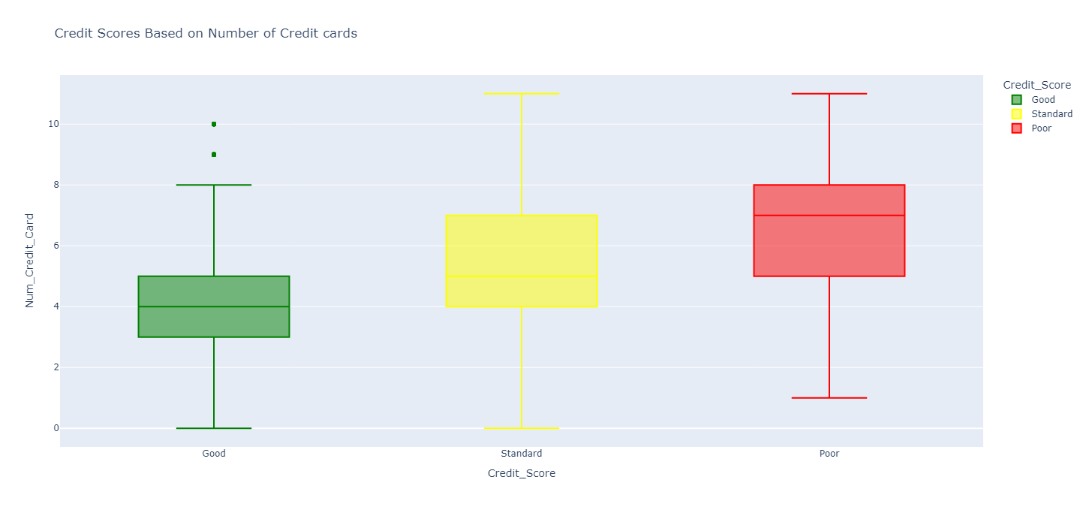
* Credit Scores Based on Number of Bank Accounts:

Box Plot: Illustrates the relationship between the number of bank accounts and credit scores, shedding light on financial stability indicators.

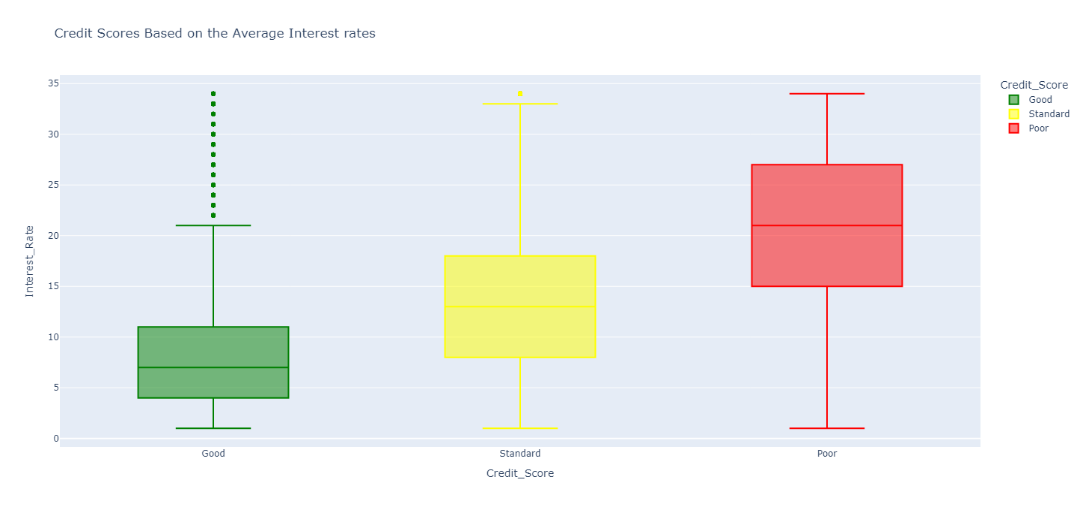


* Credit Scores Based on Number of Credit Cards:

Box Plot: Analyzes the effect of the number of credit cards on credit scores, highlighting how credit management influences credit ratings.



* Credit Scores Based on Interest Rates:

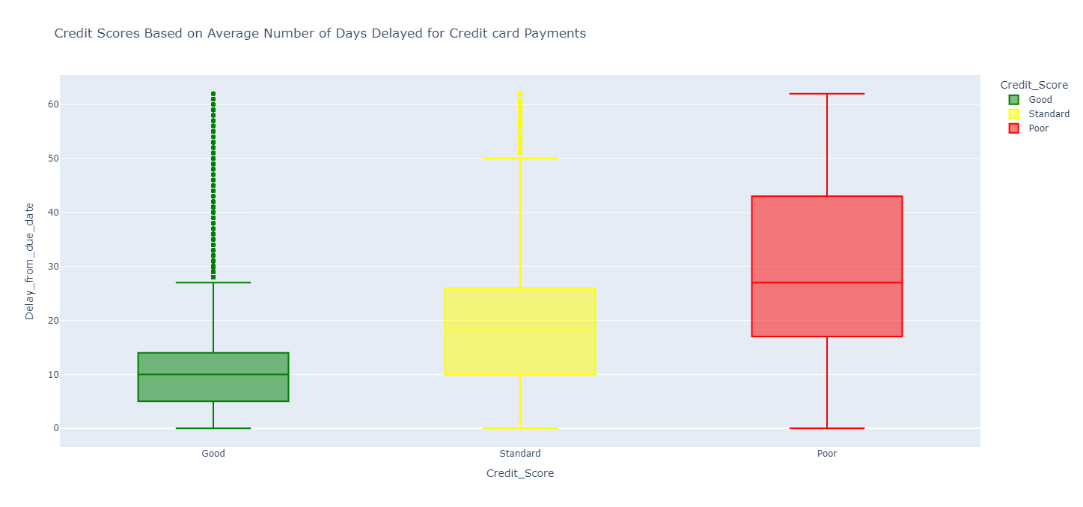
Box Plot: Displays how average interest rates affect credit scores, revealing the impact of financial costs on creditworthiness.

* Credit Scores Based on Number of Loans:

Box Plot: Provides insights into how the number of loans influences credit scores, indicating financial strain or stability.

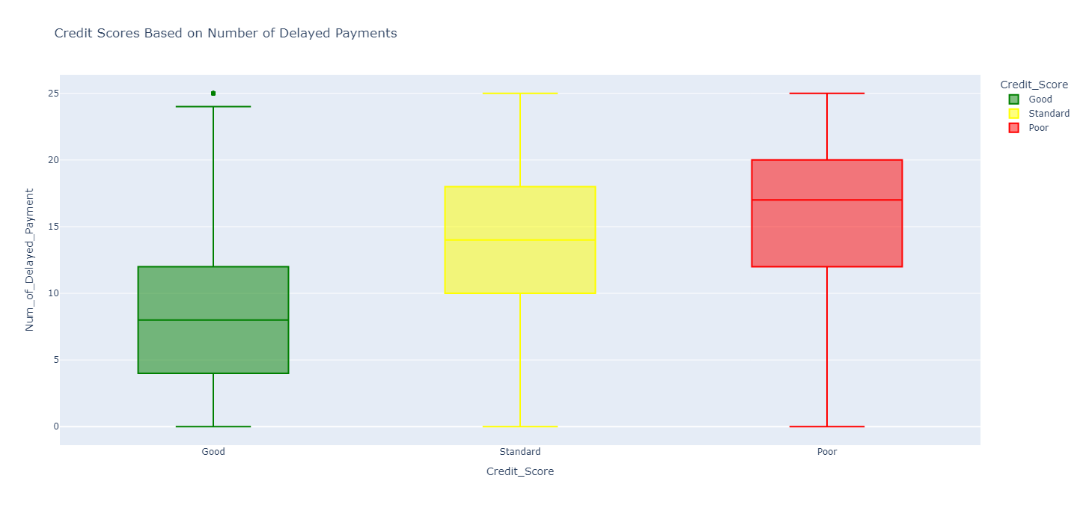
* Credit Scores Based on Delay from Due Date:

Box Plot: Shows the relationship between payment delays and credit scores, emphasizing the importance of timely payments.

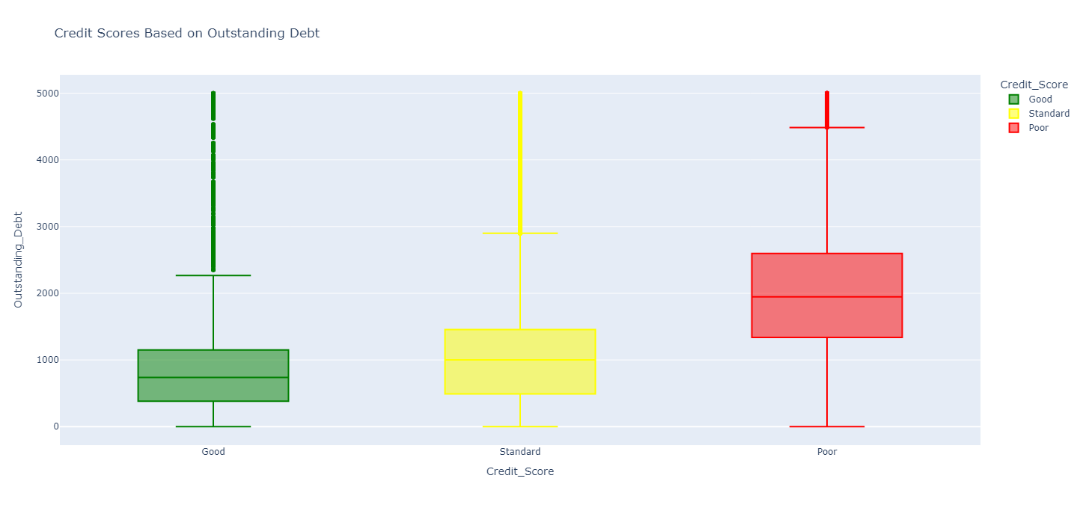


* Credit Scores Based on Number of Delayed Payments:

Box Plot: Examines how the frequency of delayed payments affects credit scores, highlighting the impact of payment history on creditworthiness.

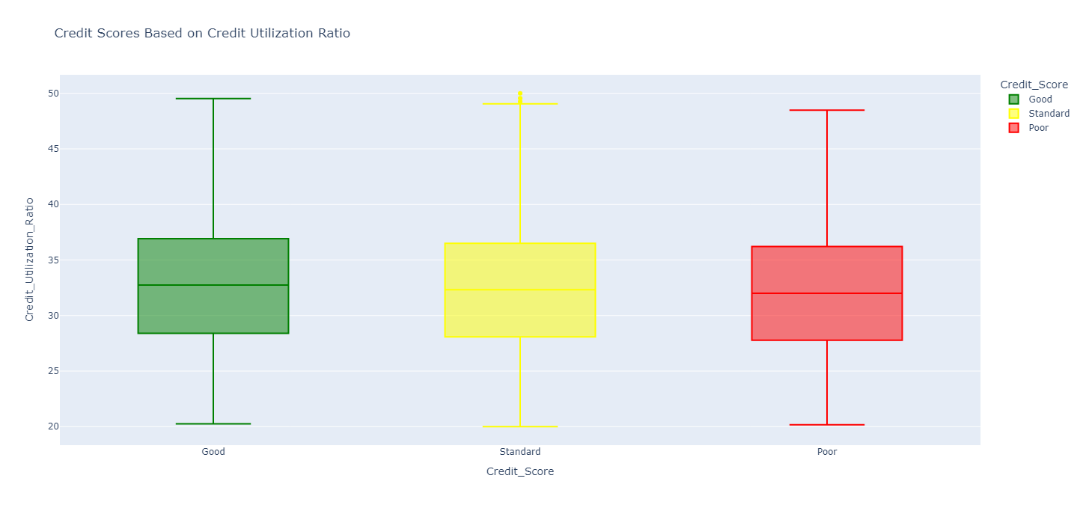


* Credit Scores Based on Outstanding Debt:

Box Plot: Visualizes the effect of outstanding debt on credit scores, indicating financial overextension and its correlation with credit ratings.

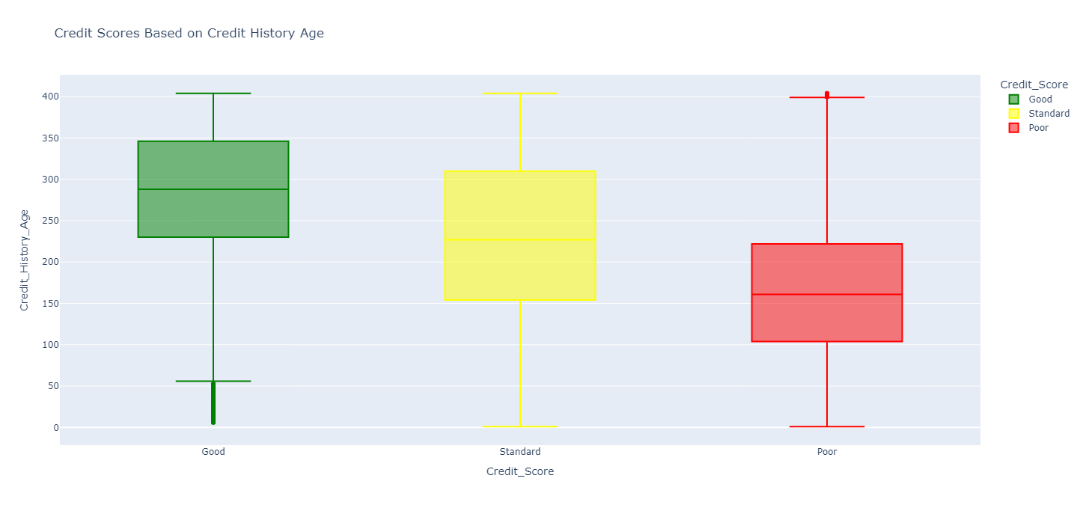
* Credit Scores Based on Credit Utilization Ratio:

Box Plot: Shows how credit utilization impacts credit scores, with lower utilization generally correlating with higher scores.

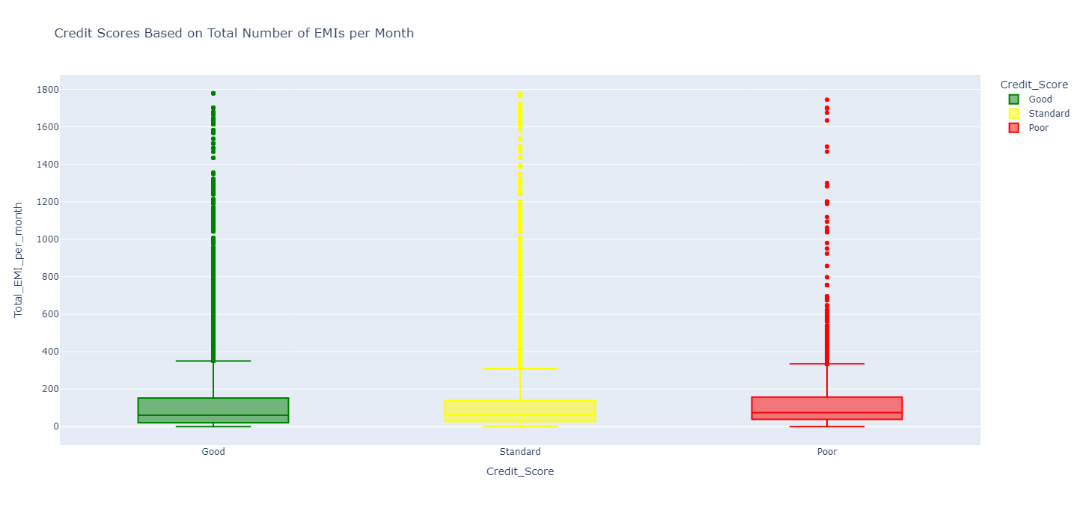


* Credit Scores Based on Credit History Age:

Box Plot: Analyzes the relationship between the length of credit history and credit scores, with longer histories often reflecting better creditworthiness.

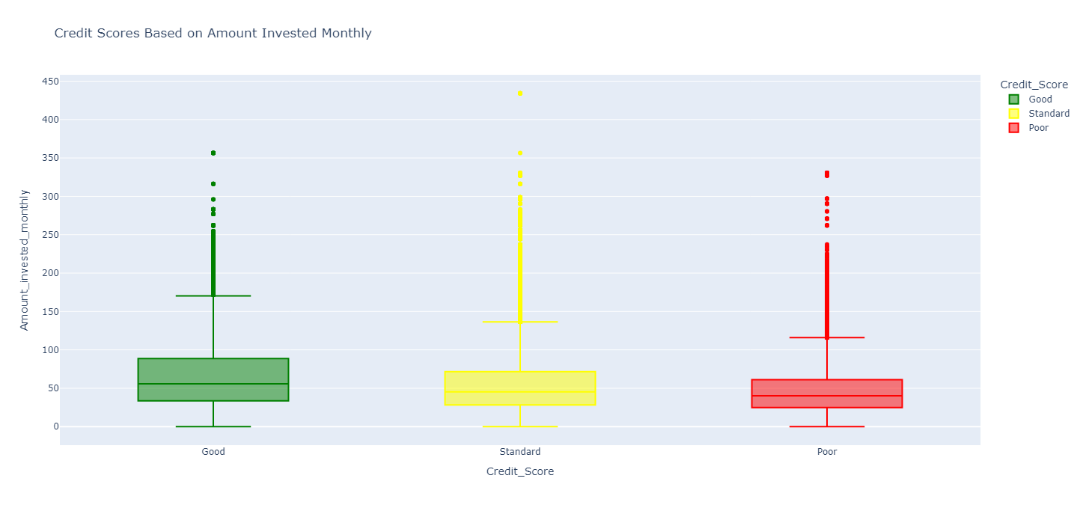


* Credit Scores Based on Total Number of EMIs per Month:

Box Plot: Examines how managing multiple EMIs affects credit scores, highlighting the financial burden and its impact.

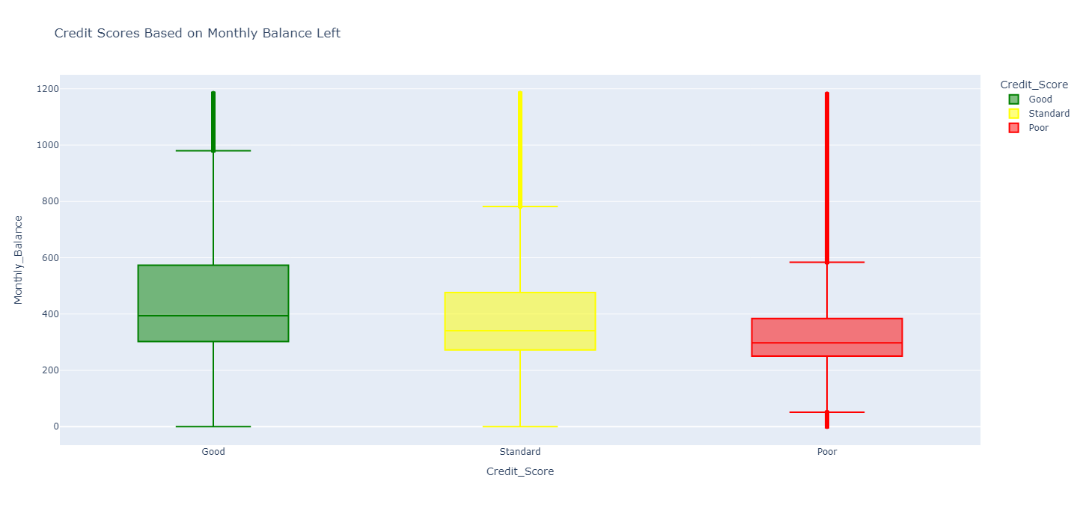
* Credit Scores Based on Amount Invested Monthly:

Box Plot: Provides insights into how monthly investment amounts correlate with credit scores, indicating financial planning habits.



* Credit Scores Based on Monthly Balance: Box Plot:

Shows how the remaining balance after expenses affects credit scores, reflecting overall financial health.



1. Model Performance and Evaluation

The Random Forest Classifier was evaluated on the test set, achieving an overall accuracy of 84.14%. This metric indicates that the model correctly classified credit scores 84.14% of the time. To understand the performance of the model in more detail, we examined precision, recall, and F1 scores for each class, as well as the confusion matrix.

Classification Report:

Class 0 (Poor Credit Score):

Precision: 0.81

Recall: 0.82

F1-Score: 0.81

Support: 3527

The precision of 0.81 means that 81% of the predicted Poor credit scores were correctly classified. The recall of 0.82 indicates that 82% of actual Poor credit scores were correctly identified by the model. The F1-score of 0.81 is the harmonic mean of precision and recall, providing a balanced measure of the model’s performance for this class.

Class 1 (Standard Credit Score):

Precision: 0.82

Recall: 0.87

F1-Score: 0.85

Support: 5874

For Standard credit scores, the precision is 0.82, meaning 82% of the predicted Standard scores were correct. The recall is 0.87, indicating that 87% of actual Standard scores were captured by the model. The F1-score of 0.85 reflects the model’s strong performance in this category.

Class 2 (Good Credit Score):

Precision: 0.87

Recall: 0.83

F1-Score: 0.85

Support: 10599

The Good credit score class shows a precision of 0.87, with 87% of predictions being accurate. The recall is 0.83, meaning the model correctly identified 83% of actual good credit scores. The F1-score of 0.85 highlights the model’s effectiveness in predicting this class.

Confusion Matrix

Class 0 (Poor Credit Score): Out of 3527 instances, 2878 were correctly classified as Poor, 8 were misclassified as Standard, and 641 were incorrectly classified as Good.

Class 1 (Standard Credit Score): Out of 5874 instances, 5137 were correctly classified as Standard, 11 were misclassified as Poor, and 726 were classified as Good.

Class 2 (Good Credit Score): Out of 10599 instances, 8813 were correctly classified as Good, 655 were misclassified as Poor, and 1131 were misclassified as Standard.

1. Discussion

The results indicate that the Random Forest Classifier performs well in predicting credit scores, with a high overall accuracy and balanced performance across different classes. The model shows the highest precision and recall for the good credit score class, suggesting it is particularly effective at identifying individuals with high credit ratings. Conversely, the model has slightly lower performance for the Poor credit score class, with more misclassifications into the good category, which could be attributed to the imbalanced nature of the dataset or the complexity of distinguishing between Poor and Good credit scores.

The precision, recall, and F1 scores provide a detailed view of the model's ability to classify each credit score category, and the confusion matrix offers insight into where the model is making errors. Future improvements could focus on addressing the misclassification issues, possibly by exploring advanced techniques such as hyperparameter tuning or incorporating additional features.

1. **Conclusion**

This research successfully develops a machine learning model for credit score classification, leveraging the power of ensemble learning through a Random Forest Classifier. The model's high accuracy and detailed analysis of feature importance provide valuable insights into credit scoring, making it a useful tool for financial institutions. Future work could explore more sophisticated models or hybrid approaches to further enhance predictive performance.