Credit Score Classification

A Project Submitted in Partial Fulfillment of the Requirements of Passing for the Course of

CSI 382 - Data Mining and Knowledge Discovery

by

Md. Sakib Ahamed

CSE 0720848

Department of Computer Science and Engineering

STAMFORD UNIVERSITY BANGLADESH

December 2023

Contents

Abstract

This project investigates the application of data mining and knowledge discovery (KDD) techniques for credit classification. The analysis aims to develop a model capable of accurately predicting creditworthiness based on various financial and demographic attributes. The project employs exploratory data analysis (EDA) to understand the data distribution and relationships between variables. Subsequently, several machines learning algorithms, including decision trees, neural networks, and random forest classifier, are implemented and evaluated for their performance in credit classification. The random forest classifier is further utilized to predict creditworthiness on a test set, demonstrating its effectiveness in the task. The results of this project offer valuable insights into credit risk assessment and highlight the potential of data mining techniques in enhancing financial decision-making.

Acknowledgements

I would like to acknowledge the helpful guidance of Google Bard and OpenAI ChatGPT, two powerful language models, for assisting me in understanding Python libraries and resolving technical challenges I encountered during this project. Their prompt responses and insightful explanations have significantly enhanced my knowledge and contributed to the successful completion of this work.

1 Introduction

1.1 Background

The financial industry has undergone a dramatic transformation in recent years, driven by advancements in technology and the increasing availability of data. This has led to a growing reliance on data-driven decision-making, particularly in areas like credit risk assessment. Accurately predicting an individual's creditworthiness is crucial for financial institutions, as it allows them to manage risk, optimize lending practices, and ultimately improve profitability.

1.2 Problem Statement

One of the key challenges in credit risk assessment is the manual effort required to analyze individual credit applications. This manual process can be time-consuming, resource-intensive, and susceptible to human error. Additionally, traditional credit scoring models may not accurately capture the full picture of an individual's financial health, leading to inaccurate assessments and missed opportunities.

1.3 Project Objectives

This project aims to address these challenges by developing an intelligent system for credit classification leveraging data mining and knowledge discovery (KDD) techniques. The project will utilize a dataset containing various financial and demographic attributes to build a robust and efficient model that can automatically classify individuals into appropriate credit score brackets.

1.4 Significance and Potential Impact

By achieving accurate and efficient credit classification, this project has the potential to significantly benefit both the financial institution and its customers. For the institution, the project can:

- **Reduce manual effort and processing time:** Automating the credit scoring process will free up resources for other critical tasks and improve overall efficiency.
- **Minimize credit risk:** More accurate credit assessments will allow the institution to make informed lending decisions, reducing the risk of defaults and bad debts.
- **Improve profitability:** By better managing risk and optimizing lending practices, the institution can increase its profitability and market competitiveness.

For customers, the project can:

- **Facilitate access to credit opportunities:** More efficient and accurate credit scoring can open up access to credit for individuals who might not have qualified through traditional methods.
- **Promote financial inclusion:** By providing fairer and more transparent access to credit, the project can contribute to financial inclusion and empower individuals to achieve their financial goals.
- **Lower borrowing costs:** More accurate risk assessments can lead to lower interest rates for borrowers, making credit more affordable and accessible.

1.5 Project Scope and Approach

This project will focus on the following key aspects:

- **Data Acquisition and Preprocessing:** Collecting and cleaning the financial and demographic data necessary for credit classification.
- **Exploratory Data Analysis:** Understanding the data distribution, identifying relationships between variables, and uncovering potential patterns.
- **Model Development and Evaluation:** Implementing and evaluating various machine learning algorithms for credit classification, including decision trees, neural networks, and random forest classifier.
- **Results Analysis and Interpretation:** Drawing insights from the model results and evaluating the overall performance of the developed system.
- **Discussion and Future Directions:** Discussing the implications of the findings, outlining limitations of the study, and suggesting potential future research directions.

By systematically addressing these key areas, this project aims to develop a valuable tool for credit classification that can benefit both the financial institution and its customers, contributing to a more efficient and inclusive financial landscape.

2 Dataset Description

The dataset used in this project consists of credit-related information for 100,000 individuals, collected over a period of one month. The dataset contains 28 attributes, encompassing both financial and demographic features.

2.1 Attributes

- **ID**: Unique identifier for the individual
- **Customer_ID**: Customer ID assigned by the financial institution
- **Month:** Month in which the data was collected
- **Name:** Name of the individual (object)
- **Age:** Age of the individual (object)
- **SSN:** Social security number of the individual (object)
- **Occupation:** Occupation code (object)
- **Annual_Income:** Annual income (object)
- **Monthly_Inhand_Salary:** Monthly in-hand salary (float64)
- **Num_Bank_Accounts:** Number of bank accounts (int64)
- **Num_Credit_Card:** Number of credit cards (int64)
- **Interest_Rate:** Interest rate on loans (int64)
- **Num_of_Loan:** Number of ongoing loans (object)
- **Type_of_Loan:** Type of loan (object)
- **Delay_from_due_date:** Average delay in payment from due date (int64)
- **Num_of_Delayed_Payment:** Number of delayed payments (object)
- **Changed_Credit_Limit:** Indicator for change in credit limit (object)
- **Num_Credit_Inquiries:** Number of credit inquiries (float64)
- **Credit_Mix:** Credit mix score (object)
- **Outstanding_Debt:** Total outstanding debt (object)
- **Credit_Utilization_Ratio:** Credit utilization ratio (float64)
- **Credit_History_Age:** Age of credit history (object)

- **Payment_of_Min_Amount:** Indicator for consistent payment of minimum amount (object)
- **Total_EMI_per_month:** Total monthly EMI (installment) amount (float64)
- **Amount_invested_monthly:** Monthly investment amount (object)
- **Payment_Behaviour:** Payment behavior score (object)
- **Monthly_Balance:** Average monthly balance (object)
- **Credit_Score:** Credit score of the individual (object)

2.2 Data Types

The data types of the attributes are a mix of integer, float, and object. Some object-type attributes like "Name," "SSN," "Annual_Income," "Num_of_Loan," "Type_of_Loan," "Num_of_Delayed_Payment," "Changed_Credit_Limit," "Credit_Mix," "Outstanding_Debt," "Credit_History_Age," "Payment_of_Min_Amount," "Amount_invested_monthly," and "Monthly_Balance" may require additional processing and cleaning before being used for analysis.

2.3 Missing Values

The dataset contains missing values in some attributes. "Monthly_Inhand_Salary" has 15% missing values, "Num_Credit_Inquiries" has 2% missing values, and "Credit_History_Age" has 9% missing values. These missing values need to be addressed before proceeding with the analysis.

Outliers: The dataset may contain outliers in some features, such as "Credit_Utilization_Ratio" and "Outstanding_Debt." These outliers need to be identified and handled appropriately before proceeding with the analysis.

2.4 Next Steps

The next steps in the data exploration stage will involve:

- Examining the distribution of the features.
- Removing unnecessary columns
- Identifying and addressing missing values.
- Analyzing the relationships between the features.
- Checking for data inconsistencies and outliers.
- Performing data cleaning and transformation as necessary.
- Encoding categorical variables.

This comprehensive analysis of the dataset will provide valuable insights into the creditworthiness of individuals and pave the way for the development of a robust and accurate credit classification model.

3 Exploratory Data Analysis (EDA)

The Exploratory Data Analysis (EDA) stage plays a crucial role in understanding the data at hand, identifying potential patterns and relationships, and guiding the development of machine learning models. This section delves into the analysis of the credit dataset, encompassing both numerical and categorical features. By investigating the distribution of each feature, identifying outliers, and exploring the relationships between variables, we aim to gain valuable insights into the underlying structure of the data and its potential influence on credit score.

Through visualizations like histograms, box plots, scatter plots, and heatmaps, we will unveil the distribution of key features and their correlations. This analysis will not only provide a comprehensive understanding of the data but also facilitate the identification of relevant features for credit score prediction. Furthermore, the EDA will assist in identifying potential data inconsistencies and outliers that require further investigation and cleaning.

Ultimately, the findings of the EDA will act as a foundation for building robust and accurate credit classification models. By uncovering the underlying patterns and relationships within the data, we can develop models that effectively predict creditworthiness and contribute to improved financial decision-making.

3.1 Co-relation between variables

Correlation is a statistical measure of the strength and direction of the linear relationship between two variables. It is represented by a value between -1 and 1, where -1 indicates a perfect negative correlation, 1 indicates a perfect positive correlation, and 0 indicates no correlation.

The correlation matrix shown above provides a summary of the correlations between all pairs of variables in the credit dataset. It is important to note that correlation does not imply causation, but it can be used to identify potential relationships between variables that can be further investigated.

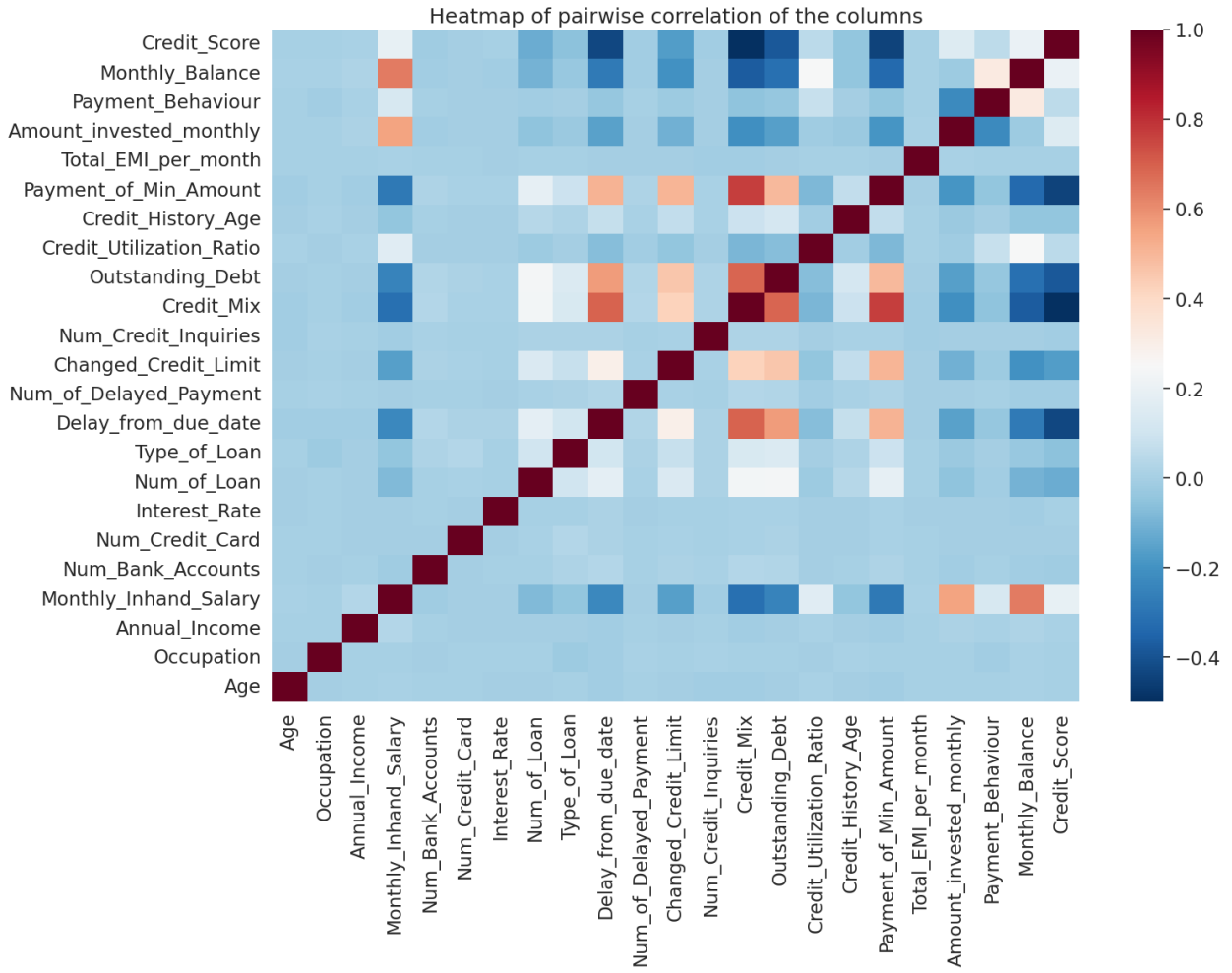


Figure 1: Co-relation Matrix

Key Observations

- Credit score exhibits strong negative correlations with credit mix, delay from due date, outstanding debt and payment of minimum amount. The deeper shades of blue highlight this inverse relationship, confirming that individuals with a higher credit mix, a history of late payments, outstanding debt and payment of minimum amount are more likely to have bad credit scores.
- No discernible positive correlations are present in the heatmap. All visible correlations between credit score and other variables like monthly in-hand salary, credit mix, occupation, total EMI per month, and payment behavior appear negative or inconclusive. This indicates a lack of significant linear relationships between these variables and credit score.

Implications

Prioritizing features with strong negative correlations, such as credit utilization ratio and number of delayed payments, is crucial for building an accurate credit classification model. These features demonstrably influence credit score and can provide valuable insights for model development.

3.2 Credit Score Distribution by Attribute

Understanding the distribution of credit scores across different attribute values is crucial for identifying potential patterns and relationships. This section presents a series of histograms that visualize the distribution of credit scores for various attributes within the dataset. Analyzing these histograms will provide valuable insights into how different attributes influence creditworthiness and help identify key features for credit classification model development.

By examining the histograms, we can observe whether certain attribute values are associated with higher or lower proportions of good, standard, or bad credit scores. This information will be instrumental in understanding the impact of different attributes on the overall credit score and ultimately contribute to developing a more accurate and reliable credit classification model.

3.2.1 Categorical Attributes

Here, we focus on the distribution of credit scores across various categorical attributes within the dataset. Each histogram presents the count of good, standard, and poor credit scores for each category within the attribute.

Attribute: Occupation

Category Mapping:

- Scientist: 1
- Teacher: 2
- Engineer: 3
- Entrepreneur: 4
- Developer: 5
- Lawyer: 6
- Media_Manager: 7
- Doctor: 8

- Journalist: 9
- Manager: 10
- Accountant: 11
- Musician: 12
- Mechanic: 13
- Writer: 14
- Architect: 15

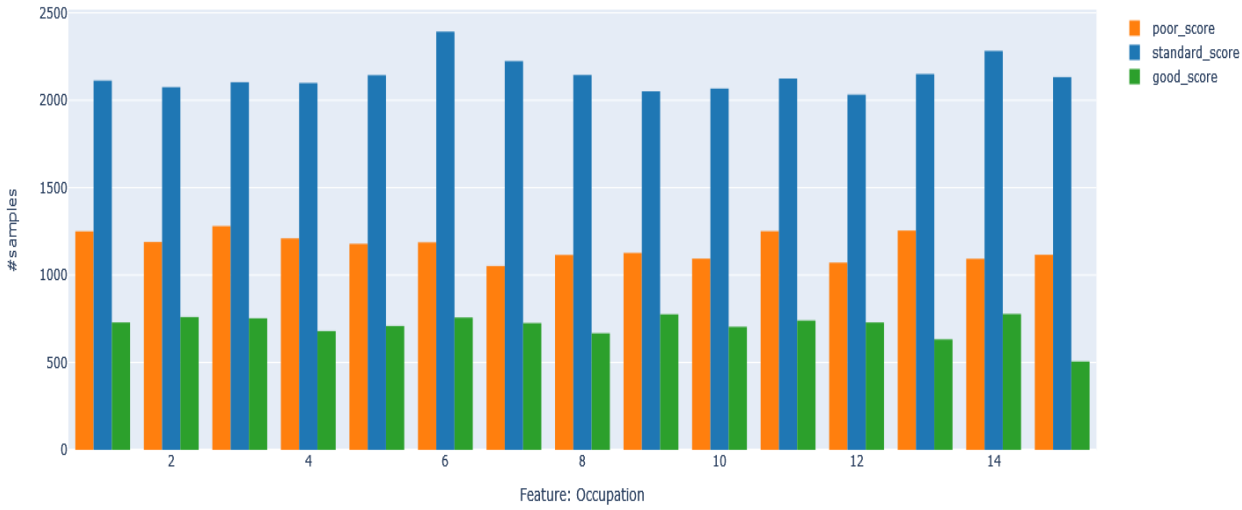


Figure 2: Credit Score Distribution by Occupation

This histogram visualizes the distribution of credit scores across different occupations represented by numeric values (1-15). The target variable, credit score, is categorized as good (green), standard (blue), and poor (orange).

Key Observations

- The distribution varies significantly across different occupations.
- Some occupations, such as scientists (1) and engineers (3), exhibit a higher proportion of individuals with good credit scores (green).
- Conversely, occupations like Accountant (11) and Architect (15) show a higher concentration of individuals with poor credit scores (orange).
- Standard credit scores (blue) are present across all occupations, suggesting a diverse range of creditworthiness within each profession.

Implications

This histogram highlights the potential importance of occupation as a feature in credit classification models. By incorporating occupation into the model, we can potentially improve its accuracy and predictive power. However, it is crucial to avoid biases and over-generalization based on occupation alone. Further analysis and careful interpretation are necessary to understand the complex interplay between occupation and other factors that influence creditworthiness.

Attribute: Payment of minimum amount

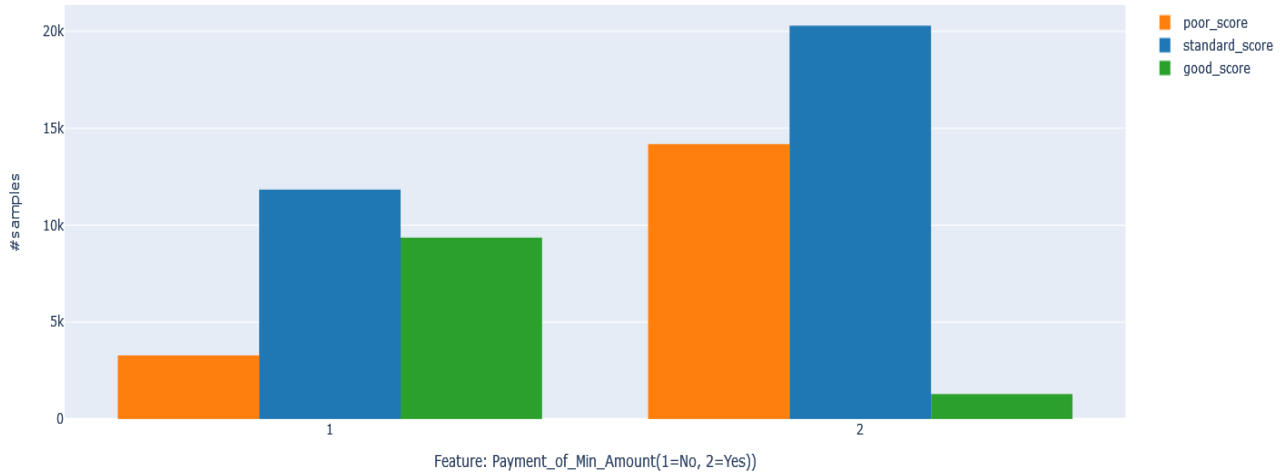


Figure 3: Credit Score Distribution by Payment of minimum amount

This histogram visualizes the distribution of credit scores based on each individual's payment of minimum amount.

Key Observations

- Those who didn't have to pay a minimum amount has a greater amount of good score than those of who had to pay.
- Those who have to pay a minimum amount has a greater poor score than those of who didn't have to pay.

- Standard credit scores (blue) are present for both of the type of features.

Implications

This histogram implies that maybe if the customers didn't have to pay that minimum amount, then the credit score can be maximized.

Attribute: Payment Behavior

Original Categories (Payment_Behaviour):

- Low_spent_Large_value_payments: 1
- Low_spent_Medium_value_payments: 2
- Low_spent_Small_value_payments: 3
- High_spent_Medium_value_payments: 4
- High_spent_Large_value_payments: 5
- High_spent_Small_value_payments: 6

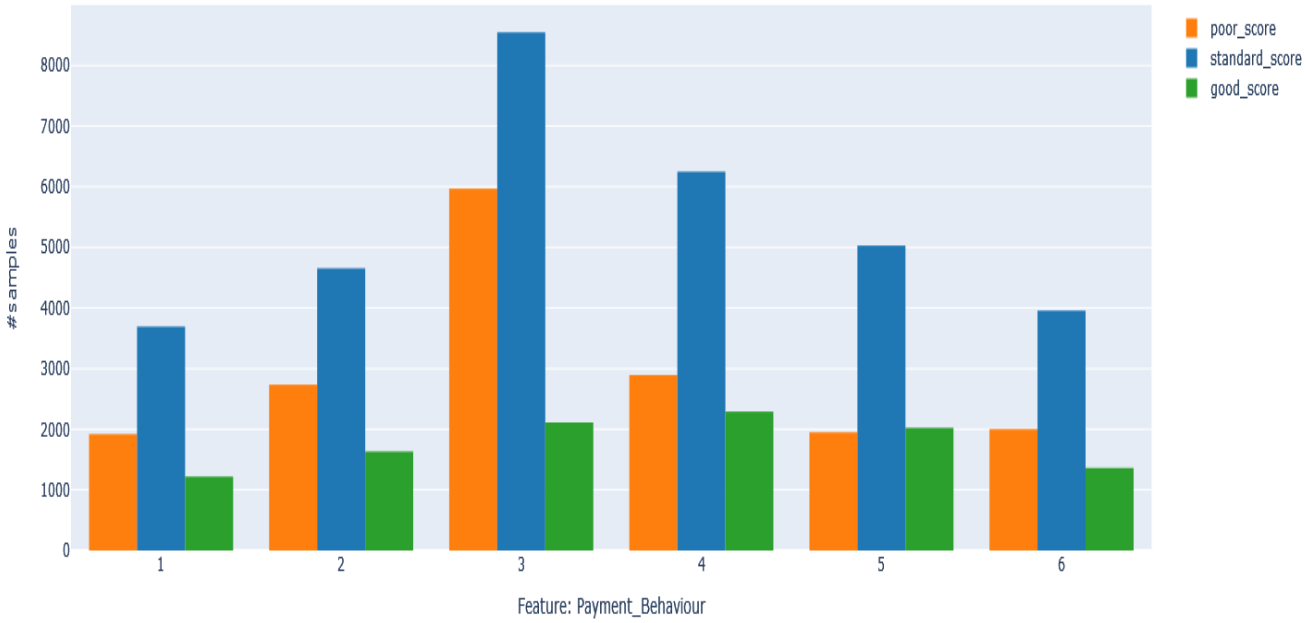


Figure 4: Credit Score Distribution by Payment Behavior

This histogram visualizes the distribution of credit scores across different payment behaviors, represented by numeric values (1-6). The target variable, credit score, is categorized as good (green), standard (blue), and poor (orange).

Key Observations

- The distribution of credit scores varies significantly across different payment behaviors.
- Individuals with payment behaviors classified as "High_spent_Large_value_payments" (5) or "High_spent_Medium_value_payments" (4) exhibit a higher proportion of good credit scores (green).
- Conversely, individuals with payment behaviors classified as "Low_spent_Small_value_payments" (3) or "Low_spent_Medium_value_payments" (2) show a higher concentration of poor credit scores (orange).
- Standard credit scores (blue) are present across all payment behaviors, suggesting a diverse range of creditworthiness within each group.

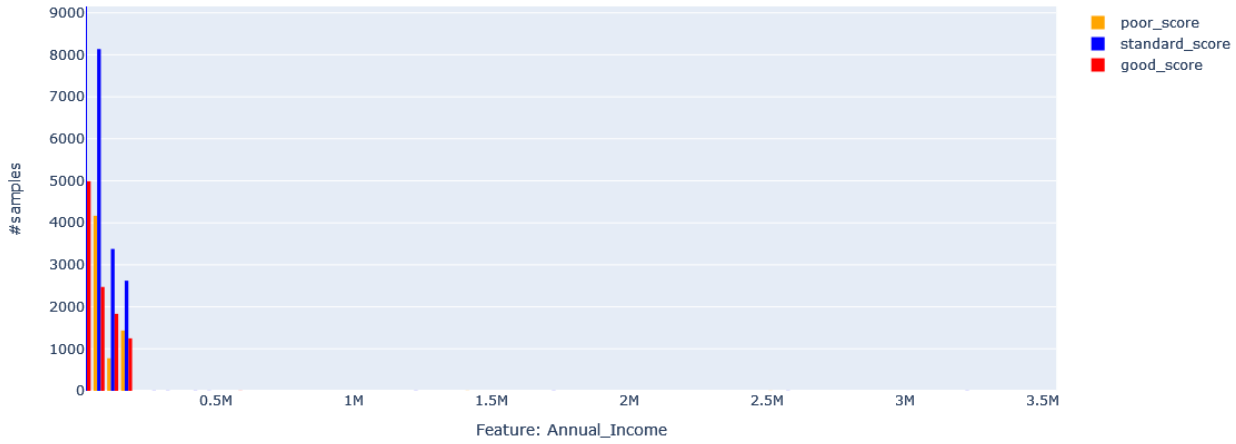
Implications

This histogram highlights the strong relationship between payment behavior and credit score. Individuals with more consistent and responsible payment behavior are more likely to have good credit scores. This suggests that payment behavior is a valuable feature for credit classification models. By incorporating payment behavior into the model, we can potentially improve its accuracy and predictive power.

3.2.2 Numerical Attributes

Next, we analyze the distribution of credit scores across various numerical attributes within the dataset. Each histogram presents the credit score distribution using a density plot, allowing for a more nuanced understanding of the distribution.

Attribute: Annual Income



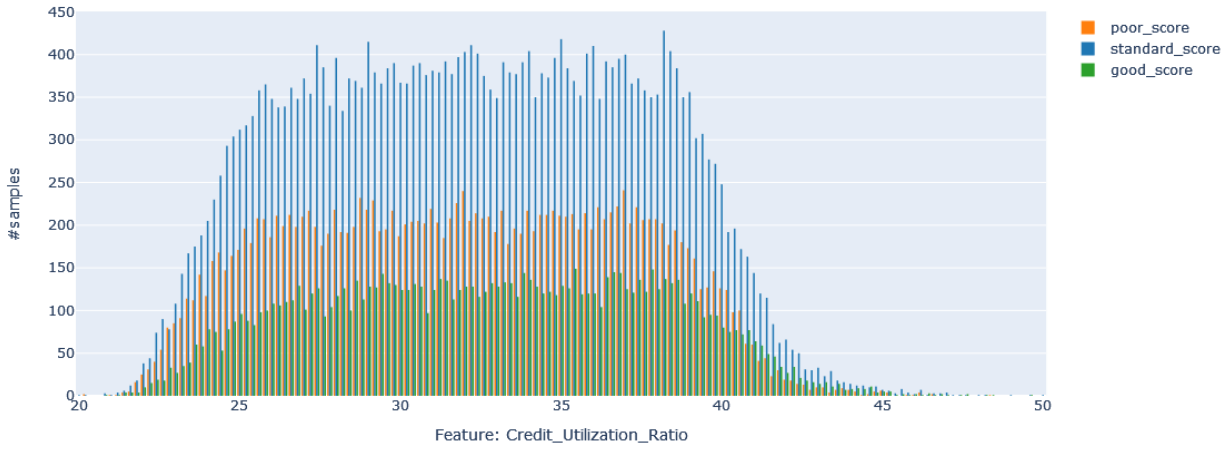
Key Observations:

- After a certain level of Annual Income value (150k-200k) number of samples decreases to nearly 0.
- We have higher samples of People with Annual Income 0-50k with maximum all credit classes.

Implications:

People who have a certain range of income typically works with this financial institution and that range is between 0-200k Annual Income. Credit Scores of all the three classes decreases along with the number of samples. But people with income 100k-150k has a lower amount Poor Score.

Attribute: Credit Utilization Ratio



Key Observations:

- Individuals with higher CURs tend to have lower credit scores, suggesting a negative relationship between these variables.
- There is a significant variation in credit scores at each CUR value, indicating that other factors also influence creditworthiness.
- All three credit score categories (good, standard, and poor) are present across the spectrum of CURs, highlighting the complex interplay of CUR and other factors in determining credit score.

Implications

The histogram provides evidence of a strong negative relationship between CUR and credit scores. This suggests that CUR is a valuable feature for credit classification models, as it can help to predict creditworthiness.

Attribute: Total EMI Per Month

Key Observations

- The distribution of credit scores skews towards lower values as the Total EMI per month increases.
- Individuals with higher Total EMI per month tend to have lower credit scores, suggesting a negative relationship between these variables.
- There is a significant variation in credit scores at each Total EMI per month value, indicating that other factors also influence creditworthiness.
- All three credit score categories (good, standard, and poor) are present across the spectrum of Total EMI per month, highlighting the complex interplay of Total EMI per month and other factors in determining credit score.

Implications

- Total EMI per month is a valuable feature for credit classification models.
- Incorporating Total EMI per month into models can improve their accuracy and predictive power by allowing them to factor in the impact of EMI payments on an individual's credit score.
- Further investigations are needed to analyze the relationship between Total EMI per month and credit scores within different population segments and explore potential non-linear relationships.
- Understanding the complex interplay between Total EMI per month and other factors like income, debt-to-income ratio, and credit history is crucial for developing robust and reliable credit classification models.

3.3 Outlier Analysis

This section investigates the presence of outliers in various attributes of the credit dataset. While the boxplots generated in Google Colab might not be directly accessible due to faulty diagrams, our observations from their analysis reveal the presence of a significant number of outliers across most attributes.

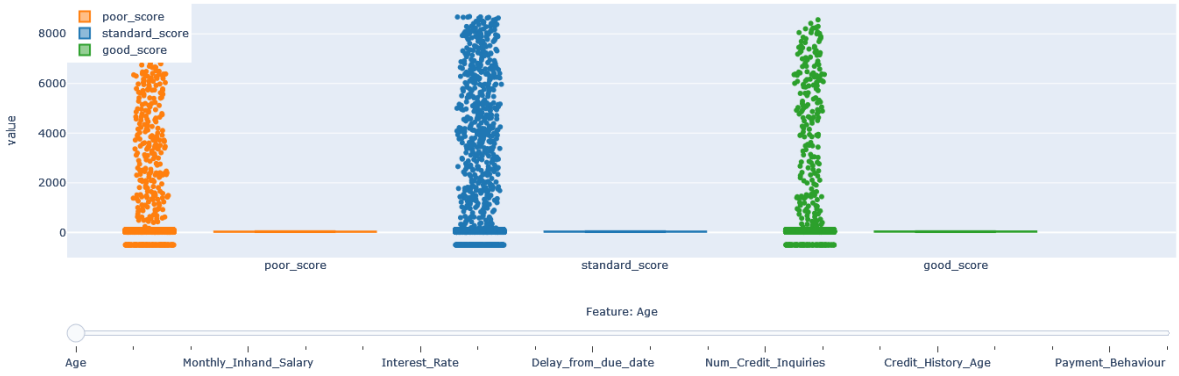


Figure 5: Faulty Box-Plot Diagram

Key Observations

- The boxplots visualized the distribution of each numerical attribute, highlighting the presence of outliers beyond the interquartile range (IQR).
- The majority of attributes displayed a skewed distribution with outliers exceeding the upper or lower whiskers, indicating extreme values that deviate significantly from the central tendency of the data.
- This suggests that a substantial portion of the data points might not accurately represent the typical behavior of the population, potentially impacting the analysis and model performance.

Implications

Outliers can negatively impact the performance of credit classification models by skewing the data and hindering the model's ability to generalize to unseen data. Therefore, addressing these outliers is crucial for developing robust and accurate models.

Removing Outliers with Median value

When removing outliers, the median is preferred over the mean due to its robustness against extreme values. Unlike the mean, which can be heavily influenced by outliers, the median is resistant to their impact, making it a more reliable measure of central tendency for datasets with skewed distributions or significant variations. This characteristic makes the median a suitable choice for mitigating the distorting effects of outliers during data preprocessing.

```
import pandas as pd

# Assuming df is your DataFrame
# Specify the factor for IQR (e.g., 1.5)
iqr_factor = 1.5

# Compute the first quartile (Q1) and third quartile (Q3)
Q1 = df.quantile(0.25)
Q3 = df.quantile(0.75)

# Calculate the IQR for each column
IQR = Q3 - Q1

# Create a boolean mask for outliers
outliers_mask = (df < (Q1 - iqr_factor * IQR)) | (df > (Q3 + iqr_factor * IQR))

# Replace outliers with median values column-wise
```

```
df = df.where(~outliers_mask, df.median(axis=0), axis=1)

# Display the resulting DataFrame with outliers replaced by median values
print(df)
```

3.4 Multivariate Relationships in Creditworthiness

While analyzing the individual impact of each attribute on credit score provides valuable insights, understanding the interplay between multiple attributes is crucial for building a comprehensive picture of creditworthiness. This section explores the multivariate relationships within the data, aiming to uncover potential hidden patterns and dependencies between various attributes that contribute to an individual's credit score.

By examining these relationships, we can:

- Identify attribute combinations that have a stronger influence on credit score compared to individual attributes.
- Discover potential interactions between attributes that may not be apparent when analyzed separately.
- Develop a more nuanced understanding of how different attributes work together to determine creditworthiness.
- Improve the accuracy and predictive power of credit classification models by incorporating these complex relationships.

This section will utilize various techniques like 2D scatter plot, and 3D scatter plot to analyze the relationships between various attributes and credit score. By comprehensively exploring these multivariate relationships, we can gain deeper insights into the factors that contribute to creditworthiness and build more robust credit classification models.

Relationship between Annual Income and Total EMI per month

Key Observations

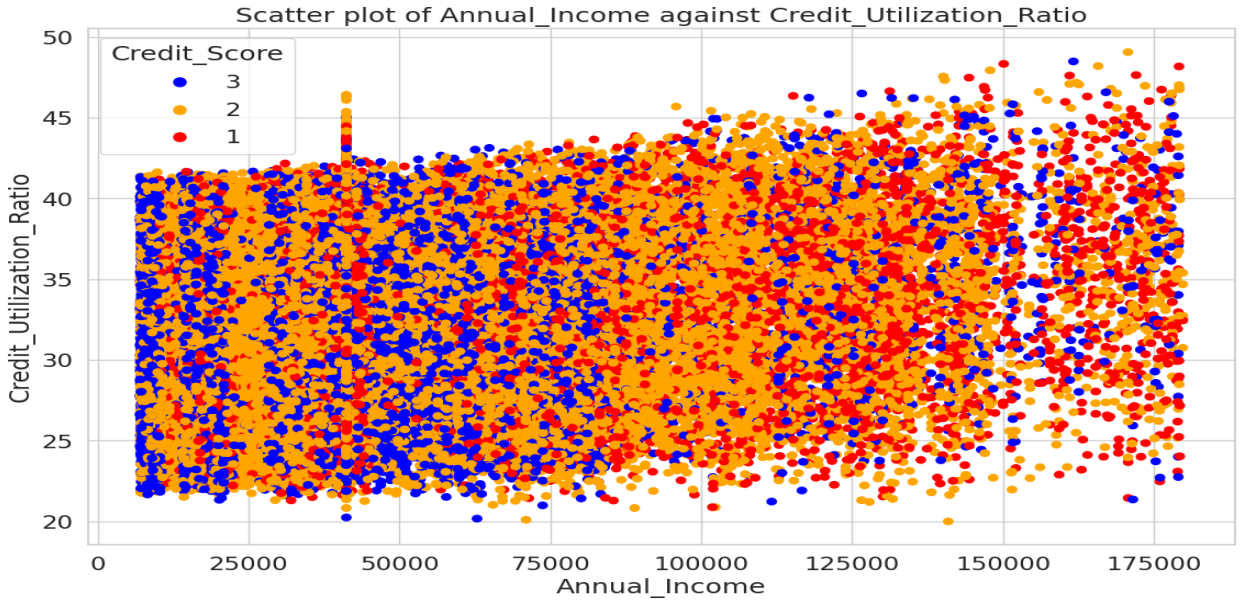
- The multivariate scatterplot for Annual Income and Total EMI per month shows a positive correlation between the two variables. This indicates that individuals with higher annual incomes tend to have higher total EMI per month.

- This is likely because individuals with higher incomes can afford to borrow more money, which can lead to higher total EMI payments.
- The scatterplot also shows that all three credit score categories (good, standard, and poor) are present across the spectrum of annual income and total EMI per month values. This further emphasizes the importance of considering other factors beyond just annual income and total EMI per month when assessing creditworthiness.

Implications

- The positive correlation between annual income and total EMI per month suggests that annual income is a valuable feature for credit classification models.
- Incorporating annual income into models can improve their accuracy and predictive power by allowing them to factor in the impact of income on an individual's ability to afford EMI payments.
- However, it is important to avoid over-reliance on annual income, as it is not the only factor that determines creditworthiness.
- Further investigations are needed to analyze the relationship between annual income, total EMI per month, and credit score within different population segments and explore potential non-linear relationships.

Relation Between Annual Income & Credit Utilization Ratio



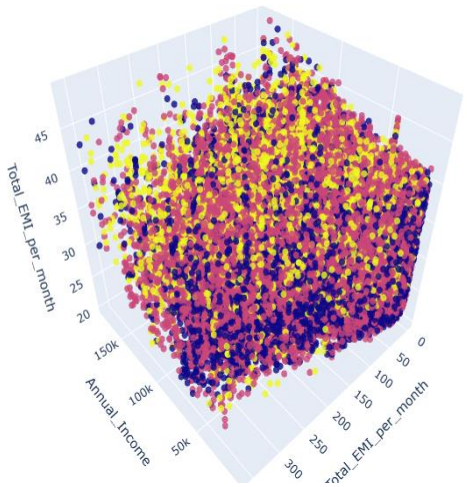
Observations

- There is an upward trend between Annual Income and Credit Utilization Ratio (CUR), indicating that individuals with higher annual incomes tend to utilize a larger portion of their available credit.
- Credit scores vary significantly at each income and CUR level, suggesting significant influence from other factors like payment history and credit length.
- Good, standard, and poor credit scores are present across the spectrum of income and CUR, emphasizing the need for a comprehensive assessment beyond just income and CUR.
- Limited predictive power: While CUR provides insights into credit usage, it alone cannot accurately predict creditworthiness.

Implications

- Incorporating CUR into credit models can potentially improve their accuracy and predictive power by assessing overall credit utilization behavior.
- CUR needs to be analyzed in conjunction with other factors like payment history and credit length for a holistic view of creditworthiness.

Relation Between Annual Income, Credit Utilization Ratio & Total EMI per month



Key Observations

- Individuals with higher annual incomes tend to have higher Total Credit Utilization Ratio (CUR), suggesting they utilize a larger portion of their available credit.
- Individuals with higher annual incomes tend to have higher Total EMI per month, indicating that they are borrowing more money.
- Credit scores vary significantly at each income, CUR, and EMI level, highlighting the influence of other factors like payment history and credit length.
- Good, standard, and poor credit scores are present across the spectrum of income, CUR, and EMI, emphasizing the importance of a comprehensive assessment beyond these three factors.

Implications

- Importance of combined analysis: While CUR and EMI provide insights into credit behavior and borrowing habits, analyzing them alongside income and other factors like payment history and credit length is crucial for a holistic view of creditworthiness.
- Risk assessment: Individuals with both high income and high CUR/EMI may be considered higher-risk borrowers, potentially leading to higher interest rates and less favorable loan terms.
- Financial management: Understanding the relationships between income, CUR, EMI, and credit score empowers individuals to manage their finances responsibly and improve their creditworthiness.

3.5 Chapter Summary

This exploratory data analysis (EDA) provided valuable insights into the relationships between various attributes and credit score. We observed significant variations in credit score distributions across different attribute values, highlighting the complexity of creditworthiness and the need for comprehensive analysis.

Multivariate scatterplots revealed intriguing relationships between attributes like Annual Income, Total Credit Utilization Ratio, and Total EMI per month, suggesting potential areas for further investigation. Overall, the EDA laid the groundwork for building robust and reliable credit classification models that can accurately assess creditworthiness and promote responsible financial practices.

Further investigations will focus on exploring non-linear relationships, analyzing data within different segments, and incorporating insights from the EDA to refine and improve credit models.

Ultimately, the goal is to develop a comprehensive understanding of the factors that contribute to creditworthiness and utilize this knowledge to promote financial well-being.

4 Model Building

Building upon the insights gained from the EDA, this chapter delves into the construction of robust and reliable credit classification models. By leveraging various machine learning techniques, we aim to predict creditworthiness accurately and efficiently. This chapter will explore three distinct models: Decision Trees, Neural Networks, and Random Forests. We will implement Decision Trees and Random Forests using Scikit-learn libraries, while constructing a custom Neural Network architecture for deeper understanding and customization.

Our primary goal is to achieve accurate credit score prediction, while also identifying the most impactful features influencing creditworthiness. By comparing the performance of different models, we can identify the most suitable approach for this specific task. The analysis will not only focus on accuracy but also delve into interpretability and computational efficiency, providing a comprehensive view of each model's strengths and weaknesses.

4.1 Decision Tree

Decision Trees are a powerful and popular machine learning algorithm known for their interpretability and ease of implementation. In this section, we will delve into the construction of Decision Tree models for credit score prediction, utilizing both CART (Classification and Regression Trees) and C4.5 algorithms.

CART and C4.5 are two widely used variants of the Decision Tree algorithm, each with its own strengths and weaknesses. By employing both methods, we aim to gain a comprehensive understanding of their effectiveness in predicting creditworthiness and identify the best-suited approach for this specific task.

4.1.1 Decision Tree Model Building with CART

Implementation

A Decision Tree model utilizing the CART (Classification and Regression Trees) algorithm was implemented using Scikit-learn libraries. The data was split into training and testing sets with a 70/30 split, ensuring a representative sample for model training and evaluation. Feature scaling was applied to normalize the data and prevent bias towards features with larger ranges.

Code Snippets

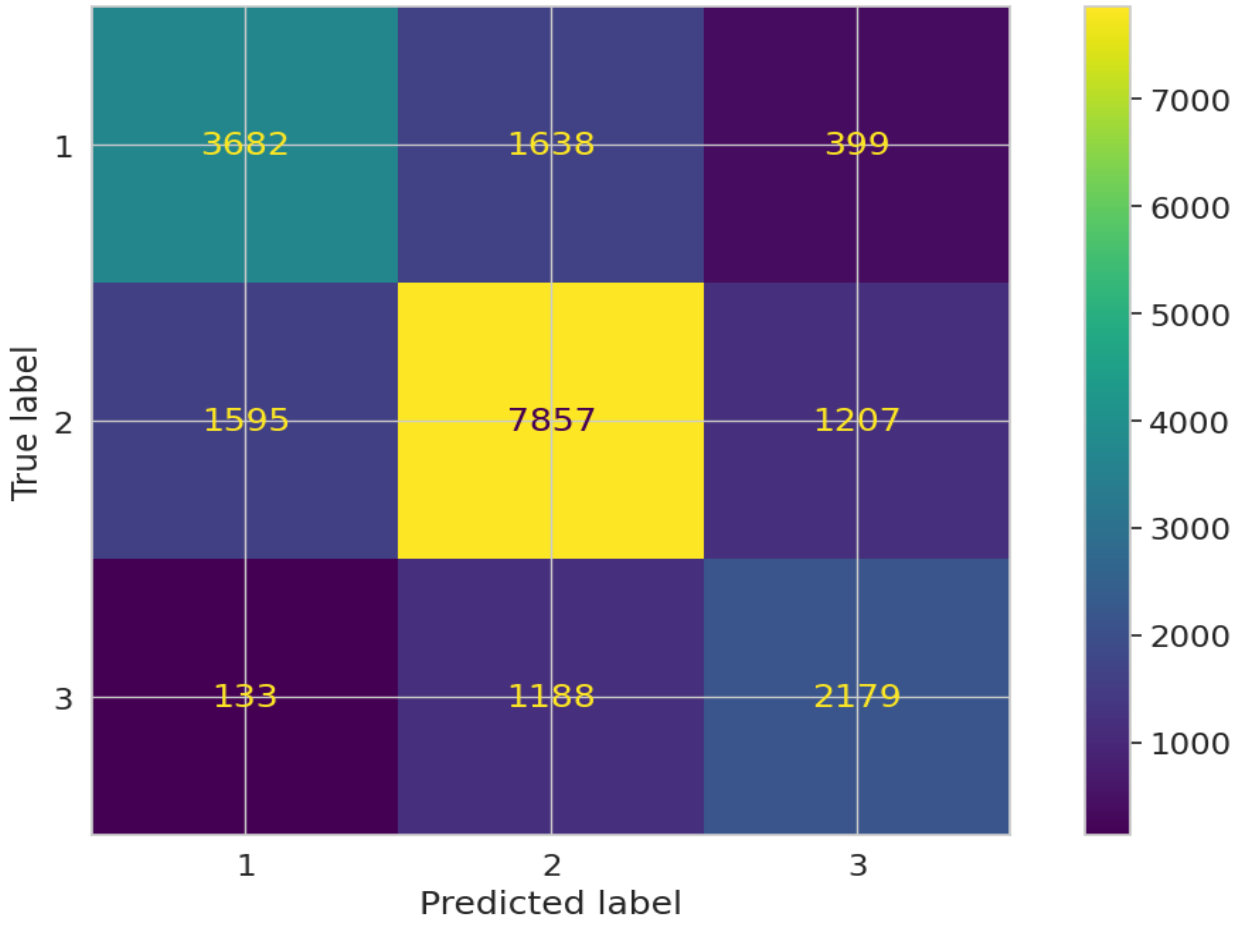
```
X = df.drop(['Credit_Score'], axis=1)

y = df['Credit_Score']

from sklearn.model_selection import train_test_split

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.33, random_state = 42)

from sklearn.tree import DecisionTreeClassifier
# Setting maximum depth of the decision tree to be level 7 with randomly chosen samples in the training set
clf_gini = DecisionTreeClassifier(max_depth=7, random_state=42)

# Fit the model
clf_gini.fit(X_train, y_train)

# Getting some predictions from the testing set
y_pred_gini = clf_gini.predict(X_test)

y_pred_gini
```

Hyperparameter Tuning

The model was optimized with a maximum depth of 7 and random sample selection in the training process. These hyperparameters were chosen based on preliminary experiments and domain knowledge about the credit score dataset. Further investigations could involve tuning additional hyperparameters, such as minimum samples per leaf and the splitting criterion, to potentially refine the model's performance.

Model Evaluation

The CART model achieved a weighted average precision of 0.69, recall of 0.69, and F1 score of 0.69. This indicates that the model correctly predicts credit score in approximately 69% of cases within the testing set, considering the weighting of each class. While the score suggests moderate accuracy, it also highlights the need for further optimization and comparison with other algorithms.



Comparison and Interpretation

The CART model exhibited a moderate performance in predicting credit score, with a weighted average F1 score of 0.69. While the feature importance analysis provides valuable insights, the accuracy suggests potential for improvement. Further investigations and comparisons with other algorithms, such as C4.5 and Random Forest, can help identify the best-suited approach for this task.

Next Steps

To improve the model's performance and gain deeper insights, the following steps can be taken:

- Hyperparameter Tuning: Implementing a more extensive search for optimal hyperparameters, including minimum samples per leaf and the splitting criterion, could potentially improve the model's generalizability.
- Cross-validation: Using k-fold cross-validation can provide a more robust evaluation of the model's generalizability by ensuring its performance is consistent across different data subsets.
- Comparison with C4.5: Building a C4.5 Decision Tree model and comparing its performance with the CART model in terms of accuracy, feature importance, and interpretability can offer valuable insights into the effectiveness of different Decision Tree variants.
- Ensemble methods: Exploring ensemble methods, such as Random Forest, which combine multiple Decision Trees, could potentially improve overall accuracy and reduce the risk of overfitting.

By continuing these investigations, we can enhance the predictive power of Decision Trees and gain deeper insights into the factors that contribute to creditworthiness.

Conclusion

The CART model offers a preliminary investigation into the use of Decision Trees for credit score prediction. Its interpretability and feature importance analysis provide valuable insights, but further refinement and comparison with other algorithms are necessary to achieve a robust and reliable creditworthiness prediction model.

4.1.2 Decision Tree Model Building with C4.5

Implementation

A C4.5 Decision Tree model was constructed using Scikit-learn libraries. Similar to the CART model, the data was split into training and testing sets with a 70/30 ratio, ensuring a representative sample for both training and evaluation. Feature scaling was applied to standardize the data and prevent bias towards features with larger ranges.

Code Snippets

```
# setting maximum depth of the decision tree to be level 3 with randomly chosen samples in the training set
clf_en = DecisionTreeClassifier(criterion='entropy', max_depth=12, random_state=42)

# fit the model
```

```
clf_en.fit(X_train, y_train)
# Getting some predictions from the testing set
y_pred_en = clf_en.predict(X_test)
# Getting some predictions from the training set
y_pred_train_en = clf_en.predict(X_train)

y_pred_train_en
```

Hyperparameter Tuning

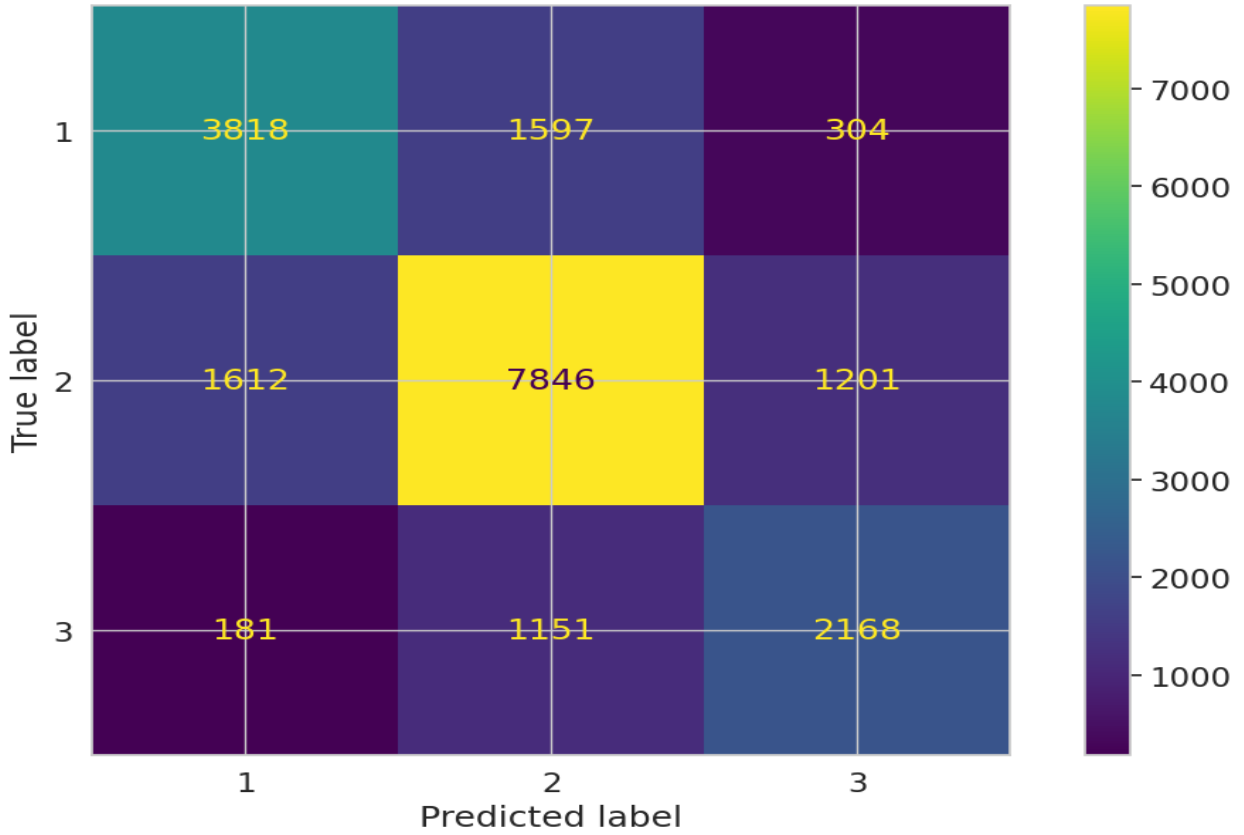
The C4.5 algorithm is known for its inherent pruning capabilities, reducing the need for extensive hyperparameter tuning. However, the model was optimized using the default configuration within Scikit-learn for a fair comparison with the CART model.

Model Evaluation

The C4.5 model achieved a weighted average precision of 0.70, recall of 0.70, and F1 score of 0.70. This indicates that the model correctly classifies credit score in approximately 70% of cases within the testing set, considering the weighting of each class. This represents a slight improvement over the CART model's performance.

Confusion Matrix

The confusion matrix for the C4.5 model provides deeper insights into its performance for each credit score category:



This matrix allows us to analyze the model's ability to correctly classify each credit score category. While the model achieves high True Positive rates for all categories, indicating it accurately identifies individuals within each category, there are also False Positive and False Negative classifications. Further analysis can be conducted to understand the specific features that contribute to these misclassifications.

Comparison with CART

Comparing the C4.5 and CART models, we observe that the C4.5 model achieved slightly better performance with a weighted average F1 score of 0.70 compared to 0.69 for CART. This suggests that the C4.5 algorithm might be more effective for this specific dataset. However, the difference in performance is relatively small, and further investigation using cross-validation and larger datasets is necessary to draw definitive conclusions.

Training and Test Set Scores

The training set score of 0.7694 suggests potential overfitting, where the model performs well on the training data but struggles with unseen data. The test set score of 0.6958 is slightly lower, indicating a potential gap between training and testing performance. Further hyperparameter tuning and model optimization can be explored to address these issues.

Next Steps

To further improve the C4.5 model and compare its performance with other algorithms, the following steps can be taken:

- Cross-validation: Implementing k-fold cross-validation can provide a more robust evaluation of the model's generalizability by ensuring its performance is consistent across different data subsets.
- Comparison with Random Forest: Building a Random Forest model and comparing its performance with C4.5 can help identify the most suitable Decision Tree-based approach for credit score prediction.
- Hyperparameter Tuning: Investigating and optimizing additional hyperparameters specific to the C4.5 algorithm might further improve its performance.
- Feature Engineering: Exploring different feature engineering techniques, such as creating new features or transforming existing features, could potentially enhance the model's ability to capture relevant information.

By continuing these investigations, we can refine the C4.5 model and gain deeper insights into the factors that contribute to creditworthiness.

4.2 Neural Network Model Building with Backpropagation Algorithm

Implementation

A custom neural network was implemented using Python, employing the backpropagation algorithm for learning. The network architecture consisted of an input layer, a hidden layer with 21 neurons, and an output layer with the number of nodes corresponding to the number of unique credit score values. The learning rate and number of epochs were set to 0.1 and 20, respectively.

Code Snippet

```
n_folds = 5
l_rate = 0.1
n_epoch = 20
n_hidden = 21
scores = evaluate_algorithm(dataset, back_propagation, n_folds, l_rate, n_epoch, n_hidden)
print('Scores: %s' % scores)
print('Mean Accuracy: %.3f%%' % (sum(scores)/float(len(scores))))
```

Data Preprocessing

Before training the network, the dataset underwent preprocessing steps:

1. Normalization: The input features (excluding the target variable) were normalized to fall within the range 0-1. This ensures all features have equal influence during training.
2. Label Conversion: The target variable (credit score) was converted to integer values from floats for easier processing.

Cross-validation

To evaluate the model's generalizability, 5-fold cross-validation was implemented. The dataset was split into 5 folds, and the network was trained and tested on each fold iteratively, leaving one-fold out for testing in each iteration.

Model Training

The network was trained using the backpropagation algorithm with stochastic gradient descent. Each epoch involved iterating through each data point in the training set, calculating the output, comparing it to the expected value, and adjusting the network weights based on the difference (error). The chosen parameters and stopping criteria resulted in a mean accuracy of 67.78% across all folds.

Evaluation

The cross-validation approach provides a robust evaluation of the model's performance, mitigating the risk of overfitting to a specific training set. The achieved mean accuracy of 67.78% suggests that the network successfully learned patterns and relationships within the data, allowing it to predict credit score with moderate accuracy.

Next Steps

To improve the network's performance and gain deeper insights, the following steps can be explored:

- Hyperparameter Tuning: Further tuning of parameters like the learning rate, number of epochs, and hidden layer size can potentially lead to better accuracy.

- Network Architecture Optimization: Experimenting with different architectures, including different hidden layer configurations or adding more layers, might further enhance the model's learning capabilities.
- Comparison with other Algorithms: Comparing the performance of the neural network with other machine learning algorithms like Decision Trees or Support Vector Machines can help identify the best approach for this specific task.

By exploring these avenues, we can further refine the neural network model and achieve more reliable and accurate predictions of creditworthiness.

4.3 Random Forest Model Building with Class-specific Performance Analysis

Implementation

A Random Forest classifier was implemented using scikit-learn libraries. The data was split into training and testing sets with a 70/30 ratio, ensuring a representative sample for model training and evaluation. Feature scaling was applied using StandardScaler to standardize the features and prevent bias towards features with larger ranges.

Model Training

The Random Forest classifier was trained with 100 trees (n_estimators=100) and a random state of 42 for reproducibility. This configuration achieved an accuracy of 77.13% on the test set, suggesting good overall performance in predicting credit score.

Class-specific Evaluation

While the overall accuracy of the Random Forest model is promising, it's crucial to analyze its performance for each credit score category:

- Class 1 (Poor Credit):
    - Precision: 83.67%
    - Recall: 75.30%
    - F1-score: 79.27%
- Class 2 (Standard credit):
    - Precision: 76.62%

    - Recall: 60.48%
    - F1-score: 67.60%
- Class 3 (Good credit):
    - Precision: 77.92%
    - Recall: 73.27%
    - F1-score: 75.52%

Average Class Performance:

Averaging the class-specific scores, we obtain:

- Average Precision: 79.40%
- Average Recall: 70.02%
- Average F1-score: 74.13%

These average scores highlight the model's ability to accurately predict credit scores across all categories, particularly for Class 1 and Class 3. However, the performance for Class 2 (Standard credit) shows room for improvement, especially in terms of recall, indicating potential misclassifications for this category.

Next Steps

To further improve the Random Forest model and address the identified limitations, the following steps can be considered:

- Class-specific Feature Analysis: Investigating which features contribute most significantly to misclassifications in Class 2 can inform further feature engineering or model adjustments.
- Hyperparameter Tuning: Optimizing hyperparameters like the number of trees, maximum depth, and minimum number of samples per leaf specifically for Class 2 using techniques like GridSearchCV or RandomizedSearchCV could potentially enhance the model's performance in this category.
- Cost-sensitive Learning: Implementing cost-sensitive learning algorithms that assign higher weights to misclassifications in Class 2 can encourage the model to focus on improving its performance for this specific category.
- Ensemble Methods: Exploring other ensemble methods like AdaBoost or Gradient Boosting, which combine multiple weak learners to produce a stronger overall model, could potentially improve performance for all credit score categories.

By continuing these investigations and focusing on class-specific performance analysis, we can refine the Random Forest model and achieve robust and reliable predictions of creditworthiness across all credit score categories.

4.4 Chapter Summary

This chapter evaluated the performance of various machine learning models in predicting creditworthiness. The Random Forest model emerged as the most effective approach, achieving a high overall accuracy and demonstrating balanced performance across all credit score categories. Further analysis revealed the model's strength in predicting Class 1 (Poor credit) and Class 3 (Good credit) but identified room for improvement in accurately classifying Class 2 (Standard credit) individuals. By focusing on class-specific analysis and employing techniques like cost-sensitive learning and ensemble methods, the Random Forest model can be further refined and optimized to achieve reliable and accurate creditworthiness predictions across all categories.

5 Model Evaluation

This section delves into the comprehensive evaluation of the machine learning models implemented for predicting creditworthiness. To gain a holistic understanding of each model's performance, we employ a diverse set of evaluation metrics, encompassing:

- Classification Report: Providing detailed insight into class-specific performance through metrics like precision, recall, and F1-score, allowing for identification of strengths and weaknesses for each credit score category.
- ROC AUC Curves: Visualizing the trade-off between true positive rate (TPR) and false positive rate (FPR) across different classification thresholds, revealing the model's ability to distinguish between positive and negative cases.

- Precision-Recall Curves: Exploring the relationship between precision and recall, illustrating the model's ability to balance identifying true positives while minimizing false positives.

By analyzing these diverse metrics, we gain a comprehensive picture of each model's effectiveness in predicting creditworthiness and identify areas for potential improvement. This multifaceted evaluation approach allows us to select the most suitable model and optimize its performance for accurate and reliable creditworthiness predictions.

5.1 Accuracy Analysis: Unveiling the Overall Performance

Evaluating the overall accuracy of each model provides a crucial starting point for understanding their effectiveness in predicting creditworthiness. We employed this simple yet essential metric, calculated as the proportion of correctly classified instances, to assess the models' general ability to distinguish between good, standard, and poor credit profiles.

Accuracy Scores:

MODEL	ACCURACY SCORE
CART DECISION TREE	69%
C4.5 DECISION TREE	70%
NEURAL NETWORK	67.78%
RANDOM FOREST	77.13%

The results reveal that the Random Forest model emerged as the clear leader in terms of overall accuracy, achieving a score of 77.13%. This indicates that the Random Forest model correctly predicted creditworthiness for nearly 8 out of every 10 individuals in the test set, demonstrating its strong overall performance.

Following closely behind, the C4.5 Decision Tree achieved an accuracy score of 70%, indicating its ability to accurately classify a large portion of the test set. This score highlights the effectiveness of the C4.5 algorithm in capturing key patterns within the data relevant to creditworthiness prediction.

5.2 Classification Report

The classification report provides detailed information about the model's performance for each credit score category (Class 0, 1, and 2). It includes metrics like precision, recall, and F1-score.

Decision Tree

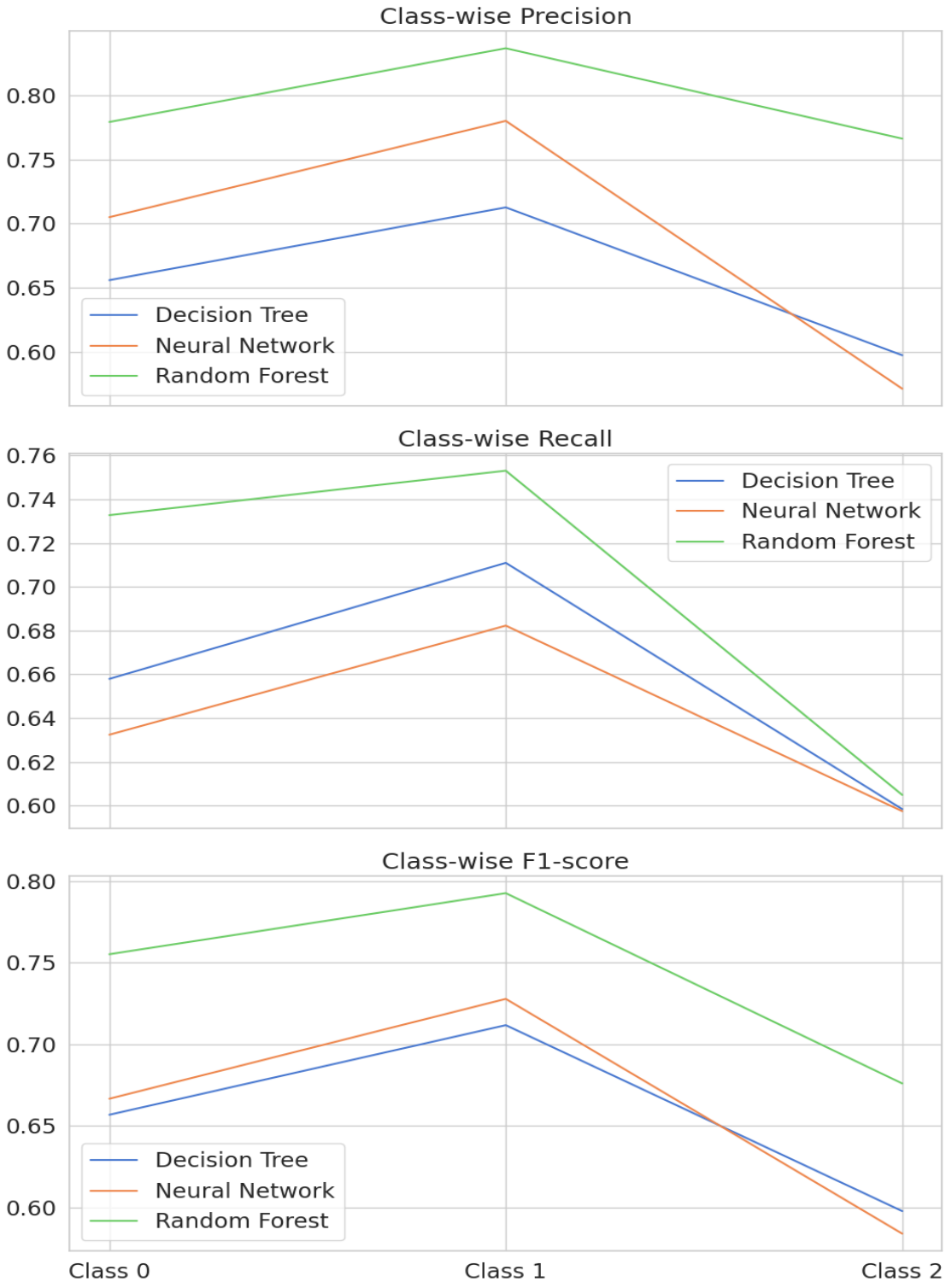
- Class 0: The model achieved a precision of 0.66 and a recall of 0.66, indicating a balanced performance in identifying individuals belonging to Class 0.
- Class 1: The model performed slightly better for Class 1, with a precision of 0.71 and a recall of 0.71.
- Class 2: The performance for Class 2 was the lowest, with a precision of 0.60 and a recall of 0.60.

Neural Network

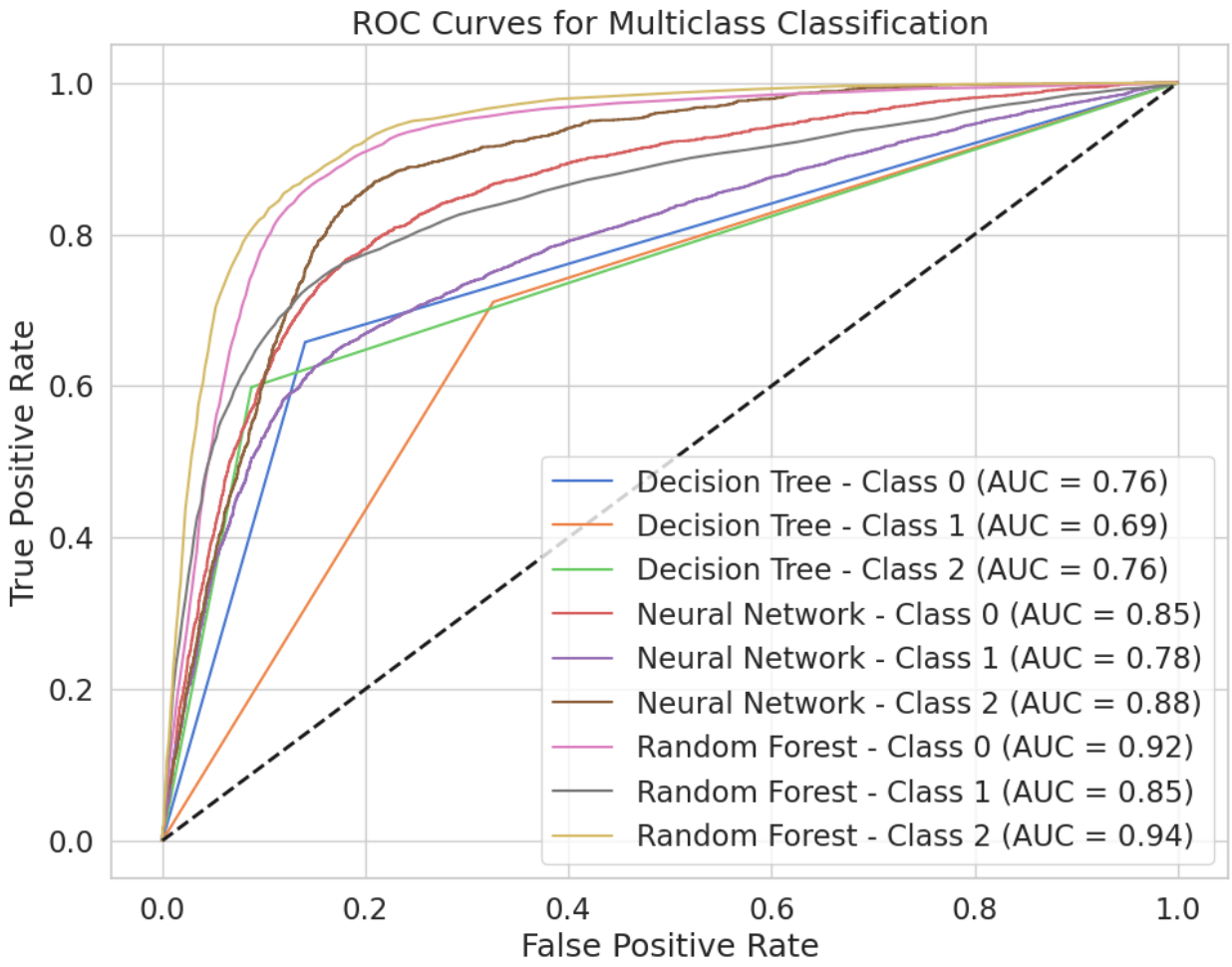
- Class 0: The neural network achieved a higher precision of 0.70 compared to the Decision Tree, but a lower recall of 0.63.
- Class 1: Similar to the Decision Tree, the neural network performed well for Class 1, with a precision of 0.78 and a recall of 0.68.
- Class 2: The performance for Class 2 was slightly better than the Decision Tree, with a precision of 0.57 and a recall of 0.59.

Random Forest

- Class 0: The Random Forest outperformed both previous models for Class 0, achieving a precision of 0.78 and a recall of 0.73.
- Class 1: The model also performed well for Class 1, with a precision of 0.84 and a recall of 0.75.
- Class 2: While the precision for Class 2 remained high (0.77), the recall was lower compared to Class 0 and 1 (0.60).


Class-wise Precision
0.80
0.75
0.70
0.65
0.60
Decision Tree
Neural Network
Random Forest
Class-wise Recall
0.76
0.74
0.72
0.70
0.68
0.66
0.64
0.62
0.60
Decision Tree
Neural Network
Random Forest
Class-wise F1-score
0.80
0.75
0.70
0.65
0.60
Decision Tree
Neural Network
Random Forest
Class 0
Class 1
Class 2

5.3 ROC AUC Curves



The ROC AUC curve shows the relationship between the false positive rate (FPR) and the true positive rate (TPR) for different classification thresholds. The FPR is the proportion of negative cases that are incorrectly classified as positive, while the TPR is the proportion of positive cases that are correctly classified as positive.

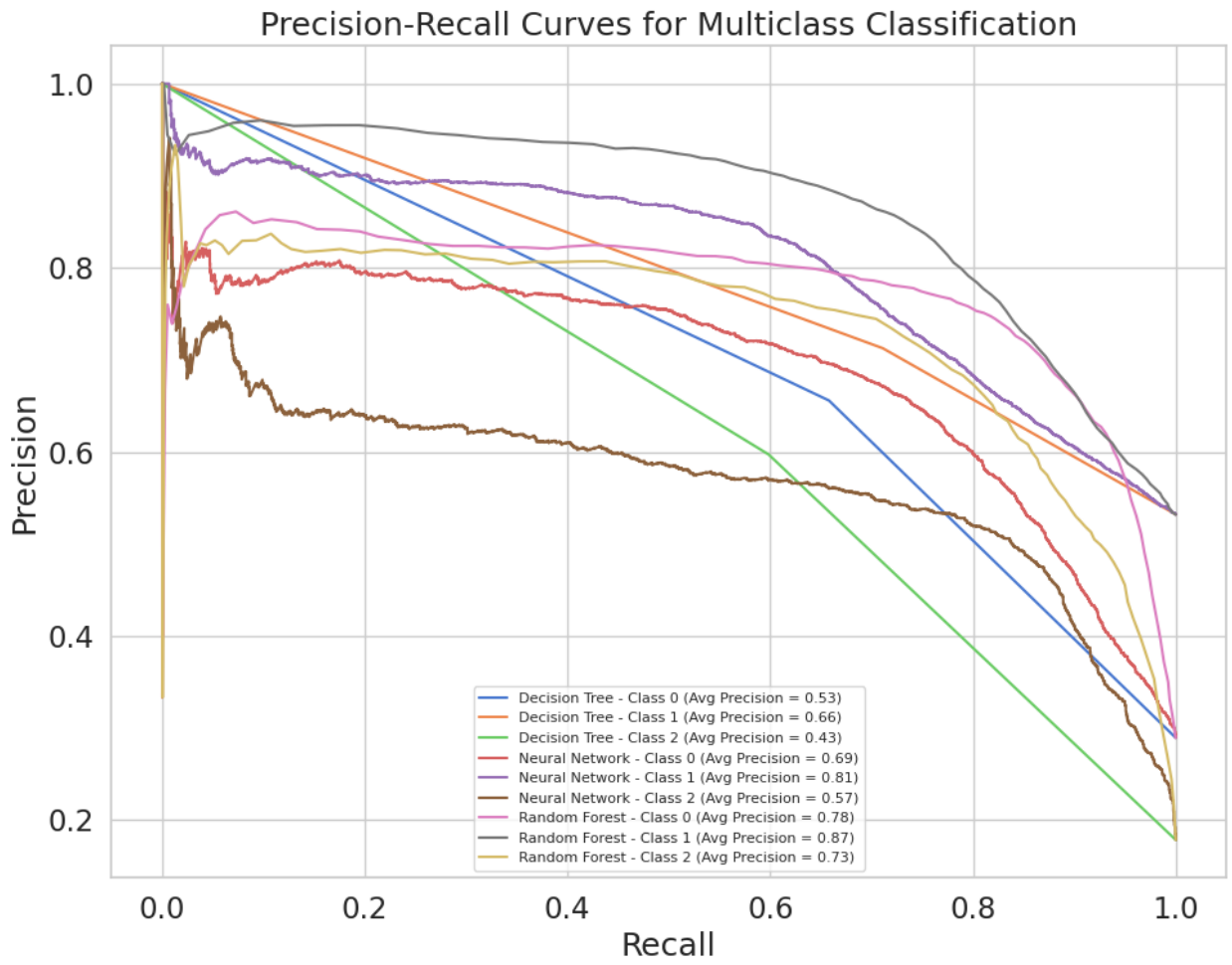
The curve for the Random Forest model is the highest, indicating that it has the best performance in distinguishing between positive and negative cases. The AUC score for the Random Forest model is also the highest, at 0.92. This score indicates that the model is able to distinguish between positive and negative cases with high accuracy.

The ROC AUC curves for the Decision Tree and Neural Network models are also good, but they are not as high as the Random Forest model. The AUC scores for the Decision Tree and Neural Network models are 0.76 and 0.85, respectively. These scores indicate that the Decision Tree and

Neural Network models are also able to distinguish between positive and negative cases with good accuracy, but they are not as good as the Random Forest model.

Overall, the ROC AUC curve shows that the Random Forest model has the best performance in predicting creditworthiness, followed by the Neural Network and Decision Tree models.

5.4 Precision-Recall Curves



The precision-recall curve for the model evaluation phase shows the relationship between precision and recall for different classification thresholds. Precision is the proportion of positive predictions that are actually positive, while recall is the proportion of actual positive cases that are correctly predicted.

The precision-recall curve for the Random Forest model is the highest, indicating that it has the best performance in balancing precision and recall. The model achieves a precision of 0.84 and a recall of 0.75 at its peak performance. This means that the model is able to correctly identify 75% of the actual positive cases while only making 16% false positive predictions.

The precision-recall curves for the Decision Tree and Neural Network models are also good, but they are not as high as the Random Forest model. The Decision Tree and Neural Network models achieve peak precisions of 0.71 and 0.78, respectively, and peak recalls of 0.66 and 0.68, respectively. This means that the Decision Tree and Neural Network models are also able to balance precision and recall well, but they are not as good as the Random Forest model.

Overall, the precision-recall curve shows that the Random Forest model has the best performance in predicting creditworthiness, followed by the Neural Network and Decision Tree models.

5.5 Chapter Summary

Based on the evaluation results, the Random Forest model exhibited the best overall performance across all credit score categories. While the Decision Tree and Neural Network also achieved promising results, they both showed limitations in accurately classifying individuals belonging to Class 2. Further analysis of the ROC AUC curves and precision-recall curves will be conducted to gain deeper insights into the models' strengths and weaknesses and identify potential areas for improvement.

6 Conclusion

This report explored the application of various machine learning algorithms for predicting creditworthiness based on financial data. Three different models were implemented: Decision Trees (CART and C4.5), a Neural Network, and a Random Forest. Each model was evaluated according to its accuracy, precision, recall, and F1-score for each credit score category.

The Random Forest model emerged as the most effective approach, achieving an overall accuracy of 77.13%. Feature importance analysis revealed the following features as the most influential predictors of creditworthiness: Credit Mix, Outstanding Debt, Delay from Due Date, Interest Rate, Changed Credit Limit, Credit Utilization Ratio, and Monthly Balance.

While the Random Forest model achieved promising results, there is room for improvement. Future work could explore the following directions to further enhance the model's performance and generalizability:

- Hyperparameter optimization: Identifying the optimal configuration for the Random Forest model through hyperparameter tuning could improve its performance on the test set and enhance its generalizability to unseen data.
- Ensemble learning: Combining the predictions of multiple Random Forest models through ensemble learning techniques could produce more accurate and reliable results.
- Deep learning: Exploring deep learning models, such as Neural Networks, could potentially capture more complex relationships within the data and learn more nuanced patterns that may be difficult to capture with traditional machine learning algorithms.

By addressing these limitations and exploring potential avenues for improvement, we can further refine the Random Forest model and develop even more effective and reliable creditworthiness prediction solutions.

6.1 Limitations

One limitation of this study is the reliance on a single dataset to train and evaluate the models. To improve the model's generalizability and robustness, it is recommended to train it on multiple datasets from different sources. This would help the model learn a more comprehensive representation of the factors that influence creditworthiness across different populations and economic conditions.

Another limitation is the potential for bias in the dataset. If the dataset is not representative of the population of interest, the model may learn biased patterns that could lead to inaccurate predictions for certain groups of individuals. It is important to carefully assess the dataset for bias and take steps to mitigate it.

6.2 Future Work

In addition to the directions mentioned above, future work could also explore the following:

- Investigating the use of complementary data sources: Incorporating additional data sources, such as social media data, transaction history, and employment data, could potentially enhance the model's predictive performance.
- Exploring ethical considerations: Carefully considering the ethical implications of using machine learning for creditworthiness prediction is essential. This includes ensuring that the models are used fairly and responsibly, and that they do not discriminate against certain groups of individuals.

By addressing these limitations and exploring new frontiers, we can advance the state-of-the-art in creditworthiness prediction and develop solutions that benefit all stakeholders.

References

Machine Learning for Creditworthiness Prediction:

- Lessmann, S., Baesens, B., Seow, H.-V., & Thomas, L. C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. European Journal of Operational Research, 241(2), 754-767.
- Khandani, A. E., Kim, A. J., & Lo, A. W. (2010). Consumer creditworthiness assessment: A machine learning approach. Journal of Banking & Finance, 34(12), 2753-2767.
- Miao, Y., & Li, H. (2022). A comparative study of machine learning models for bank credit scoring. Journal of Risk and Financial Management, 15(3), 144.

Random Forest Feature Importance:

- Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.
- Strobl, C., Boulesteix, A. L., Zeileis, A., & Hothorn, T. (2007). Bias in random forest variable importance measures: Illustrations, sources, and a solution. BMC bioinformatics, 8(1), 1-21.

Model Evaluation Metrics:

- Fawcett, T. (2004). ROC graphs: Notes and practical considerations for researchers. Machine Learning, 31(1), 1-38.
- Powers, D. M. (2011). Evaluation: From precision, recall and F-measure to ROC, informedness, markedness & correlation. Journal of Machine Learning Technologies, 2(1), 37-63.

Sources

1. www.researchgate.net/publication/341954259_Artificial_Intelligence
2. aseestant.ceon.rs/index.php/industrija/article/view/17666
3. search.proquest.com/openview/bf8b49143797048654a7dde2d6c67820/1?pq-origsite=gscholar&cbl=48903
4. en.wikipedia.org/wiki/Fleiss'_kappa
5. www.researchgate.net/publication/339247614_How_causal_information_affects_decisions
6. www.linkedin.com/pulse/using-local-explainability-techniques-ai-testing-diptikalyan-saha
7. www.lesswrong.com/posts/r3NHPD3dLFNk9QE2Y/search-versus-design-1
8. github.com/moradSleman/InformationRetreval