THE UNIVERSITY OF ARIZONA

School of Information Sciences

MASTER OF SCIENCE IN INFORMATION SCIENCE:MACHINE LEARNING

Project Report

German Bank Loan

(Machine Learning Models)

Gubbala Durga Prasanth

(Student ID: 23879524)

Abstract

The German Bank Loan project aims to develop a robust machine learning model to predict loan default based on historical customer data. Leveraging a dataset containing information such as financial indicators, credit history, loan purpose, and demographic factors, we explore the multifaceted dynamics influencing loan default predictions. The project employs a diverse set of machine learning models, including RandomForestClassifier, LogisticRegression, DecisionTreeClassifier, GradientBoostingClassifier, and KNeighborsClassifier, to comprehensively analyze and predict default scenarios. Key questions addressed include the correlation between financial indicators and default likelihood, the impact of credit history on predictions, the significance of loan purpose, and the contribution of demographic factors. This abstract provides a succinct overview of the project's objectives, methodologies, and key findings, offering valuable insights into the complexities of predicting loan defaults.

Acknowledgments

This project represents a culmination of effort and support from various sources, and I extend my sincere gratitude to those who have contributed to its realization. I would like to express my appreciation to school mentor and industrial mentor for providing valuable guidance and insights throughout the project's development. Special thanks to my batch mates who offered collaborative input and constructive feedback, enhancing the overall quality of the analysis. Additionally, I acknowledge the resources and datasets made available for this project, particularly the German Bank Loan dataset, which served as the foundation for our exploratory analysis and model development. This project would not have been possible without the collective support, and I am thankful for the opportunities and knowledge gained during its execution.

TABLE OF CONTENTS:

Introduction

The project centers around addressing a critical challenge faced by a German bank, which has been grappling with issues related to loan defaulters among its customers. The historical dataset provided contains information on various features associated with customers who have previously taken loans from the bank. In an effort to mitigate the risk of defaults, the bank aims to leverage machine learning techniques to develop a predictive model. The primary objective is to predict whether a customer is likely to default on a loan based on their historical financial behavior and other relevant attributes.

This project is motivated by the need to enhance the bank's decision-making process and minimize financial risks associated with loan defaults. By utilizing advanced machine learning algorithms, we aim to uncover patterns and relationships within the dataset that can be indicative of potential default cases. The dataset encompasses diverse aspects of customer profiles, including their financial history, employment details, and demographic information.

Key Questions:

Throughout the course of this project, several intriguing questions will be explored to gain valuable insights into the factors influencing loan default predictions. Some of the key questions include:

1. *How do different financial indicators, such as checking balance and savings balance, correlate with the likelihood of loan default?*

2. *What role does the credit history of customers play in predicting loan defaults, and how does it vary across different credit histories?*

3. *Can the purpose for which a loan is taken serve as a significant predictor of default, and are there specific loan purposes associated with higher default rates?*

4. *How do demographic factors, such as age, employment duration, and the presence of dependents, contribute to the accuracy of default predictions?*

By exploring these questions, we aim to not only build an effective predictive model for loan defaults but also gain a comprehensive understanding of the underlying factors that influence customer creditworthiness in the context of the German banking dataset.

Methods and Materials

The analyses conducted in this project are structured into two primary phases: Exploratory Data Analysis (EDA) and the development of machine learning models for predicting loan defaults.

Exploratory Data Analysis (EDA)

The initial phase of this project involved a detailed Exploratory Data Analysis (EDA) of the German Bank dataset, a comprehensive collection of historical data on customers who have availed loans. With 17 columns and 1000 rows, the dataset encompasses essential features such as employment duration, existing loans count, savings balance, percentage of income, age, and the target variable, "default" status. To provide a comprehensive overview, statistical measures such as mean, median, and standard deviation were computed for numerical features, while visualizations such as histograms and box plots were employed to understand the distribution of the data. Categorical variables were explored using bar charts, aiding in the identification of patterns. Additionally, a correlation heatmap was generated to reveal relationships between different features, shedding light on potential multicollinearity.

Following the EDA, the dataset underwent preprocessing steps, addressing missing values, encoding categorical variables, and introducing a new feature, "total_income," derived from the amount and percentage of income. This ensured that the data was prepared and devoid of inconsistencies before proceeding with model training.

Exploring the data:

Dataset Information:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 17 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   checking_balance      1000 non-null   object
 1   months_loan_duration  1000 non-null   int64
 2   credit_history        1000 non-null   object
 3   purpose               1000 non-null   object
 4   amount                1000 non-null   int64
 5   savings_balance       1000 non-null   object
 6   employment_duration   1000 non-null   object
 7   percent_of_income     1000 non-null   int64
 8   years_at_residence    1000 non-null   int64
 9   age                   1000 non-null   int64
 10  other_credit          1000 non-null   object
 11  housing               1000 non-null   object
 12  existing_loans_count  1000 non-null   int64
 13  job                   1000 non-null   object
 14  dependents            1000 non-null   int64
 15  phone                 1000 non-null   object
 16  default               1000 non-null   object
dtypes: int64(7), object(10)
memory usage: 132.9+ KB
None
```

Summary Statistics:

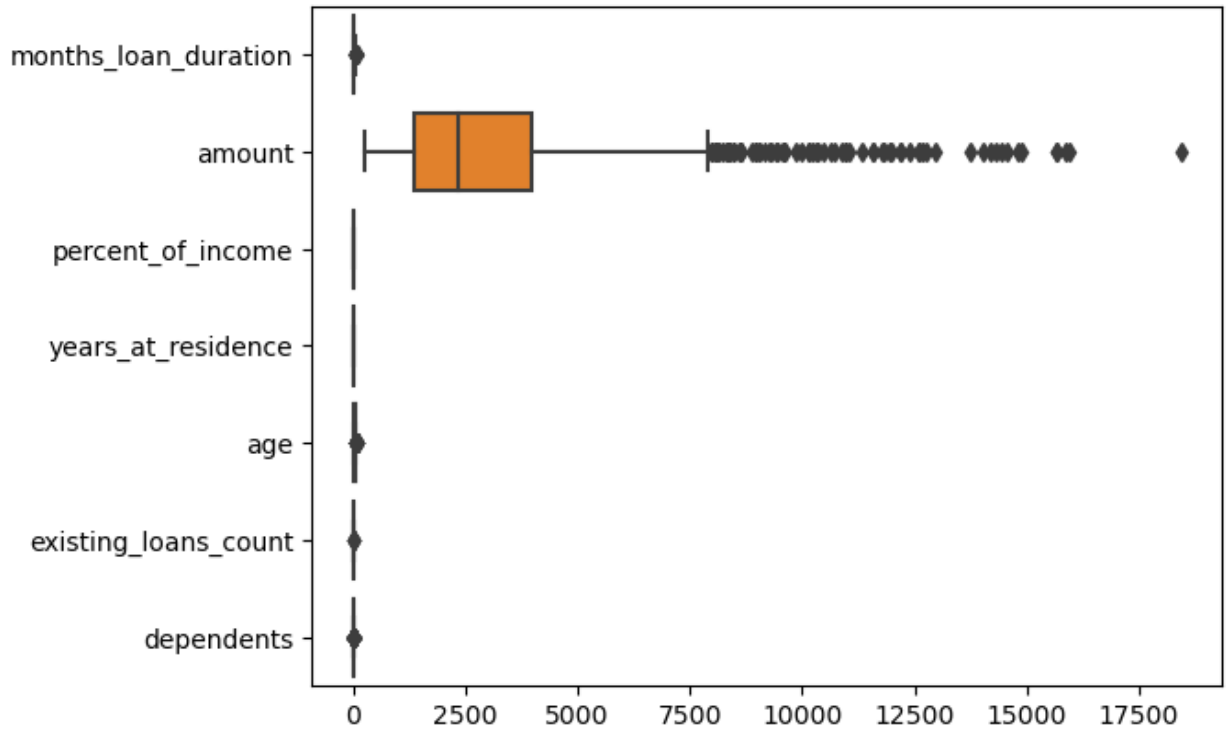
```
       months_loan_duration        amount  percent_of_income  \
count           1000.000000   1000.000000        1000.000000
mean              20.903000   3271.258000           2.973000
std               12.058814   2822.736876           1.118715
min                4.000000    250.000000           1.000000
25%               12.000000   1365.500000           2.000000
50%               18.000000   2319.500000           3.000000
75%               24.000000   3972.250000           4.000000
max               72.000000  18424.000000           4.000000

       years_at_residence          age  existing_loans_count   dependents
count         1000.000000  1000.000000           1000.000000  1000.000000
```

```
mean             2.845000    35.546000             1.407000     1.155000
std              1.103718    11.375469             0.577654     0.362086
min              1.000000    19.000000             1.000000     1.000000
25%              2.000000    27.000000             1.000000     1.000000
50%              3.000000    33.000000             1.000000     1.000000
75%              4.000000    42.000000             2.000000     1.000000
max              4.000000    75.000000             4.000000     2.000000
```

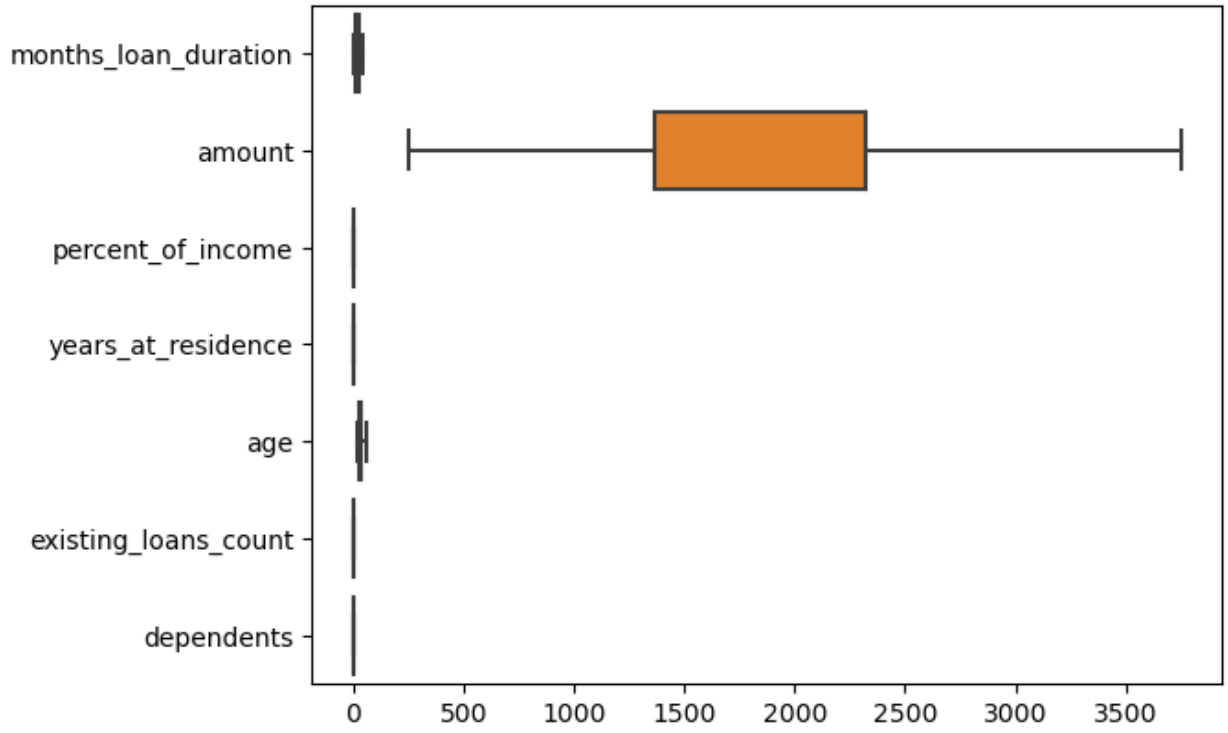
First Few Rows:

```
  checking_balance  months_loan_duration credit_history              purp
ose  \
0           < 0 DM                     6       critical  furniture/applian
ces
1       1 - 200 DM                    48           good  furniture/applian
ces
2          unknown                    12       critical             educat
ion
3           < 0 DM                    42           good  furniture/applian
ces
4           < 0 DM                    24           poor
car

   amount savings_balance employment_duration  percent_of_income  \
0    1169         unknown           > 7 years                  4
1    5951       < 100 DM         1 - 4 years                  2
2    2096       < 100 DM         4 - 7 years                  2
3    7882       < 100 DM         4 - 7 years                  2
4    4870       < 100 DM         1 - 4 years                  3

   years_at_residence  age other_credit housing  existing_loans_count  \
0                   4   67         none     own                     2
1                   2   22         none     own                     1
2                   3   49         none     own                     1
3                   4   45         none   other                     1
4                   4   53         none   other                     2

         job  dependents phone default
0    skilled           1   yes      no
1    skilled           1    no     yes
2  unskilled           2    no      no
3    skilled           2    no      no
4    skilled           2    no     yes
```

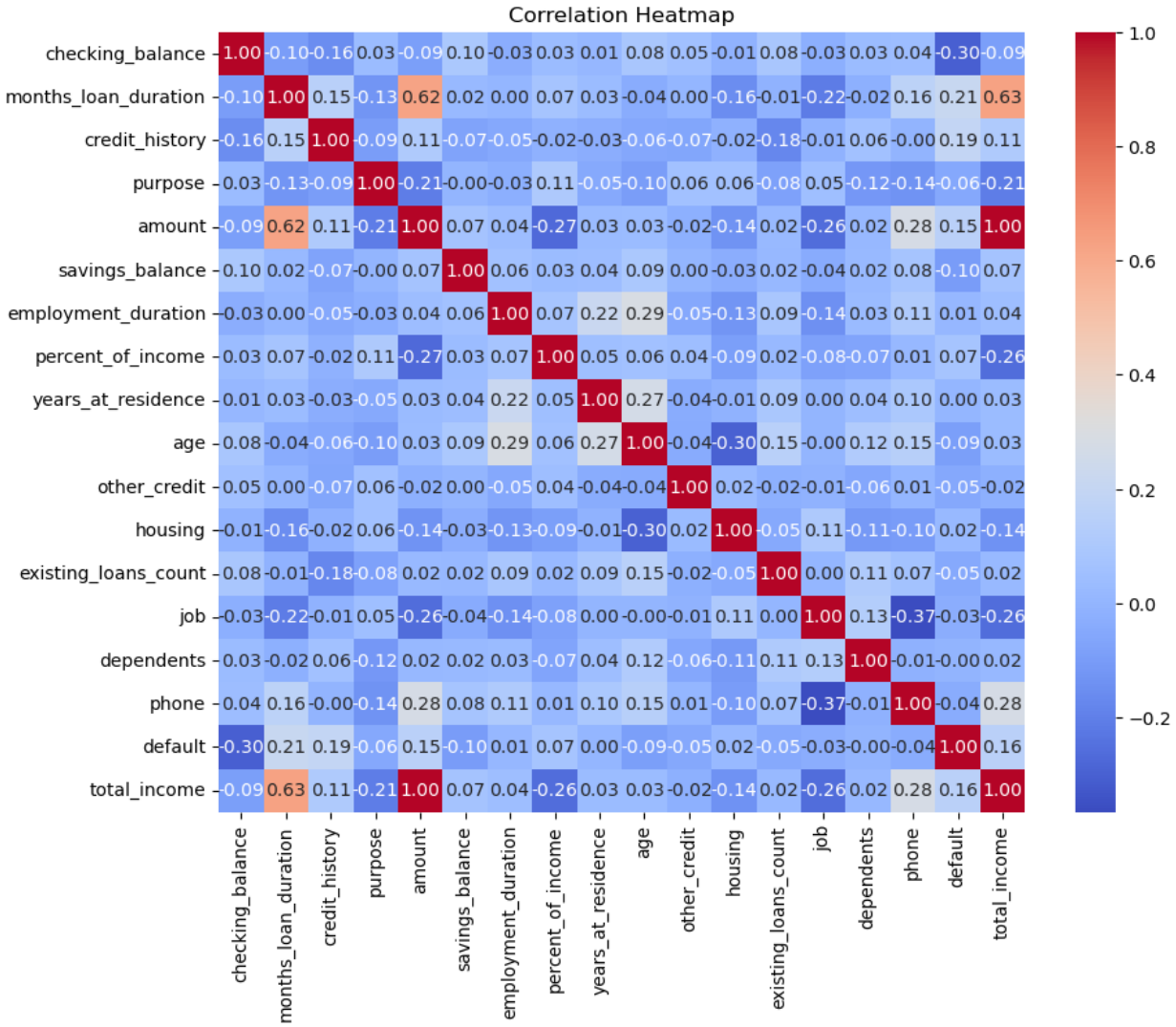
checking is there any outliers exist:



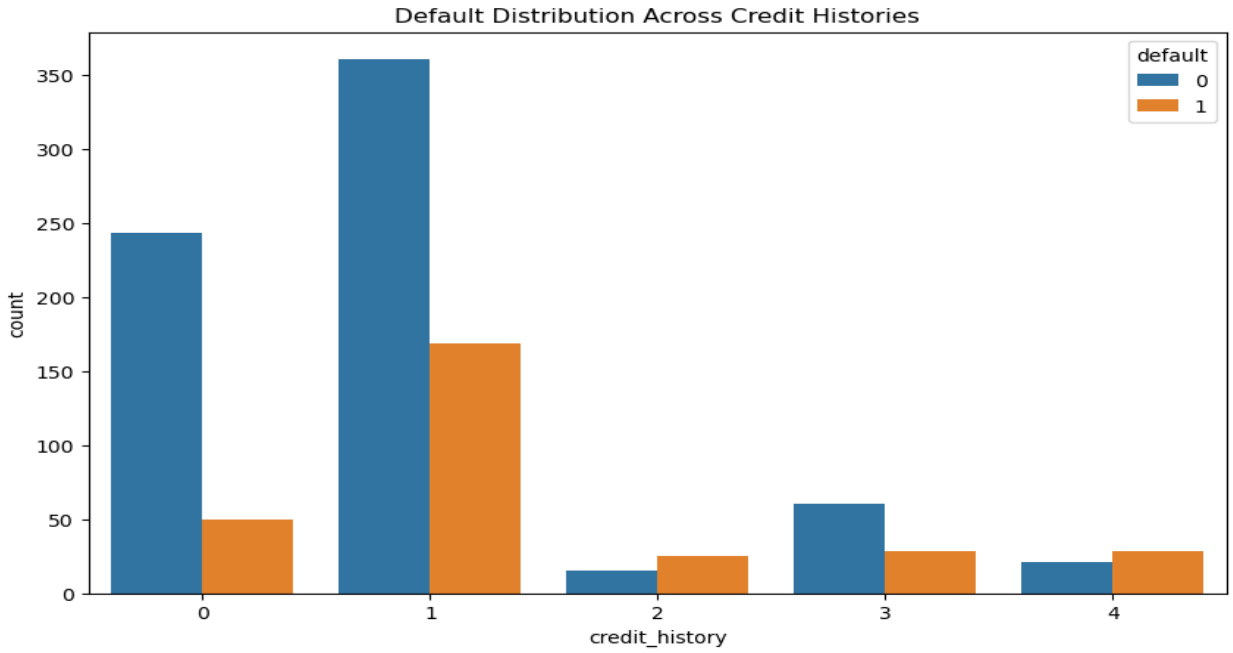
outliers handled:



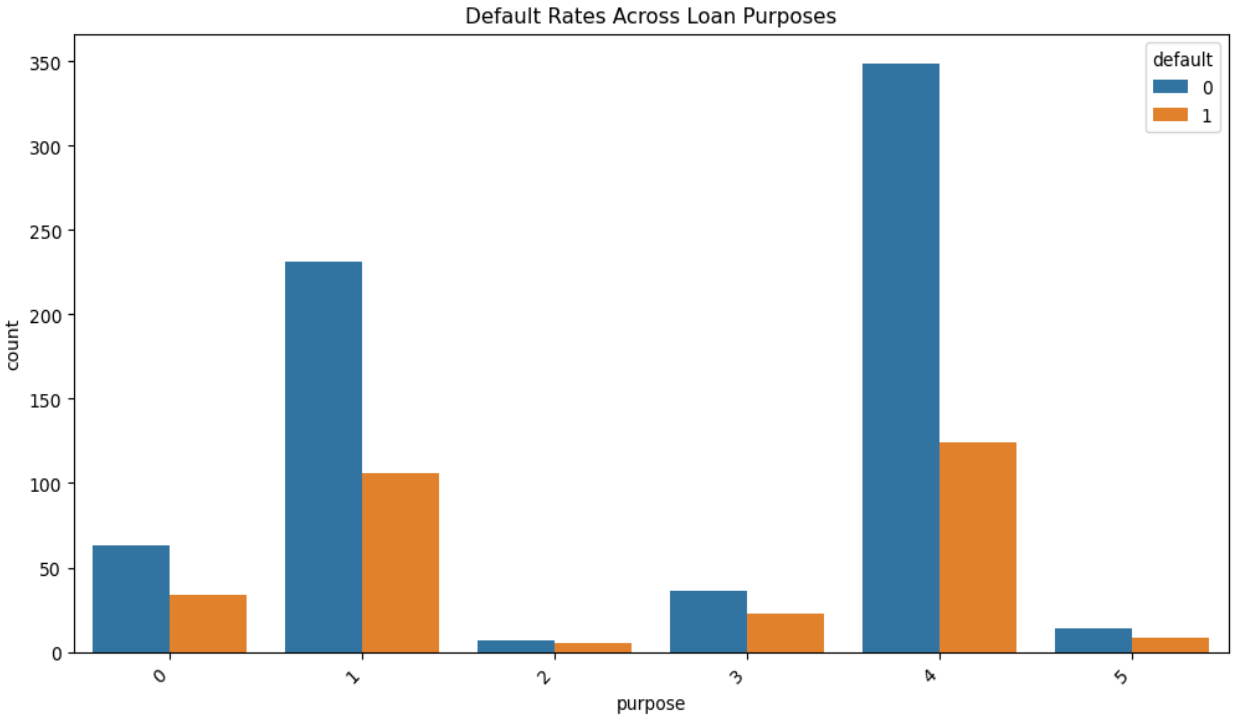
Correlation heatmaps and distribution plots:



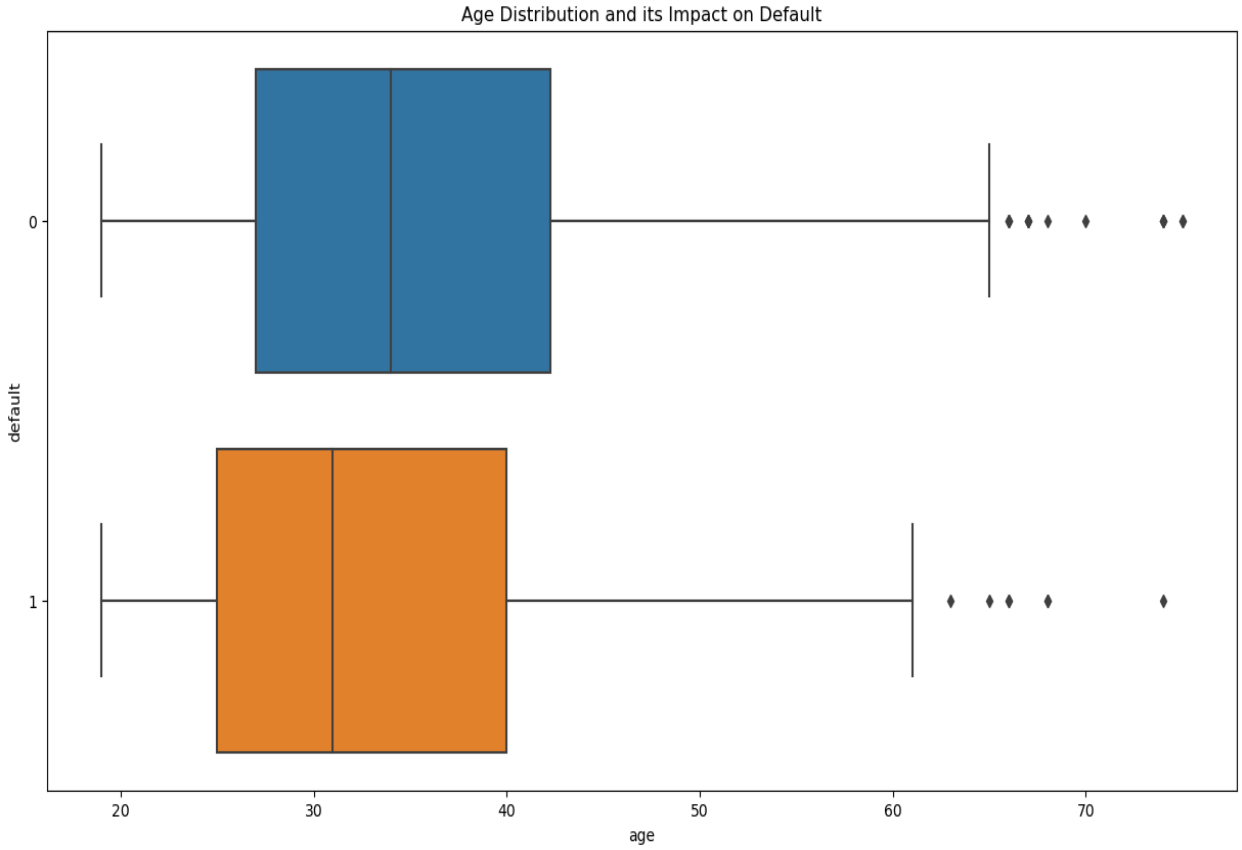
Credit History Analysis:



Loan Purpose Analysis:

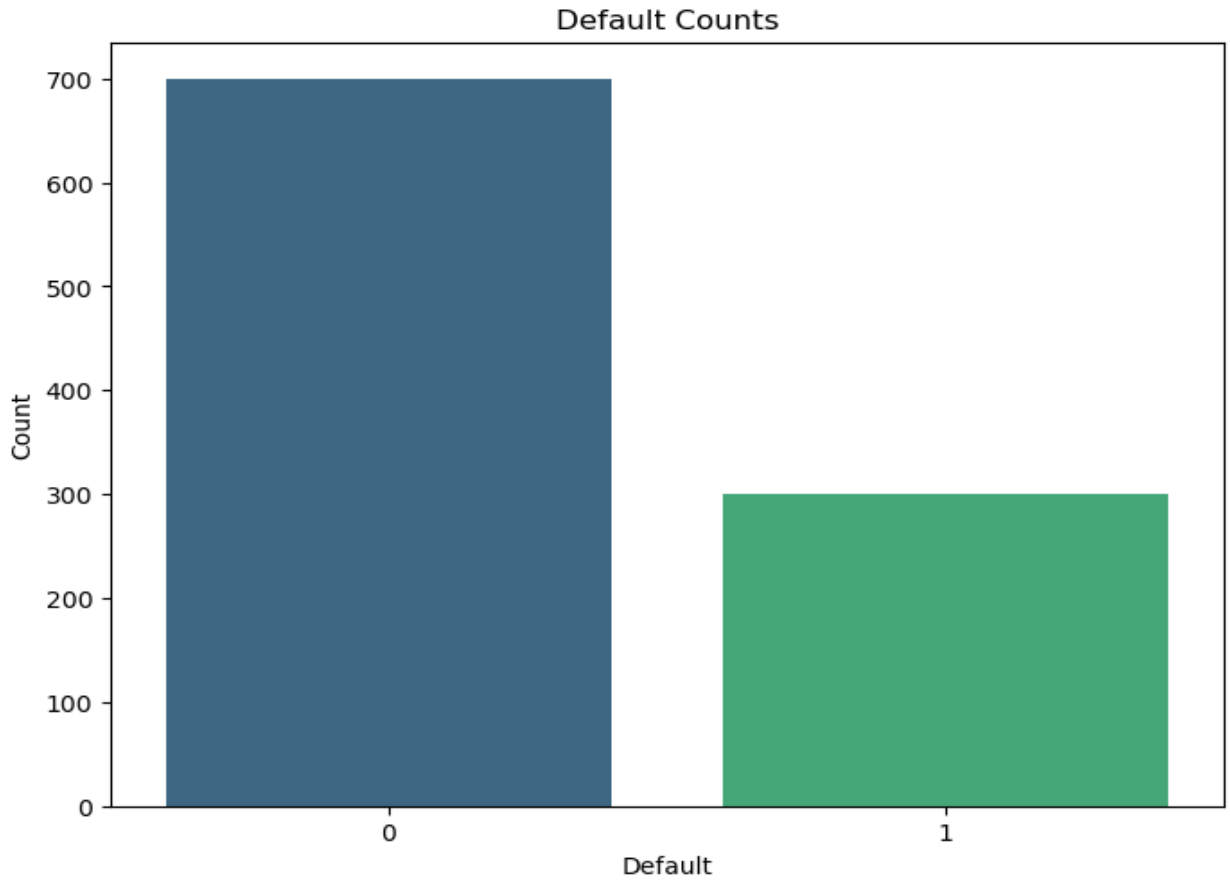


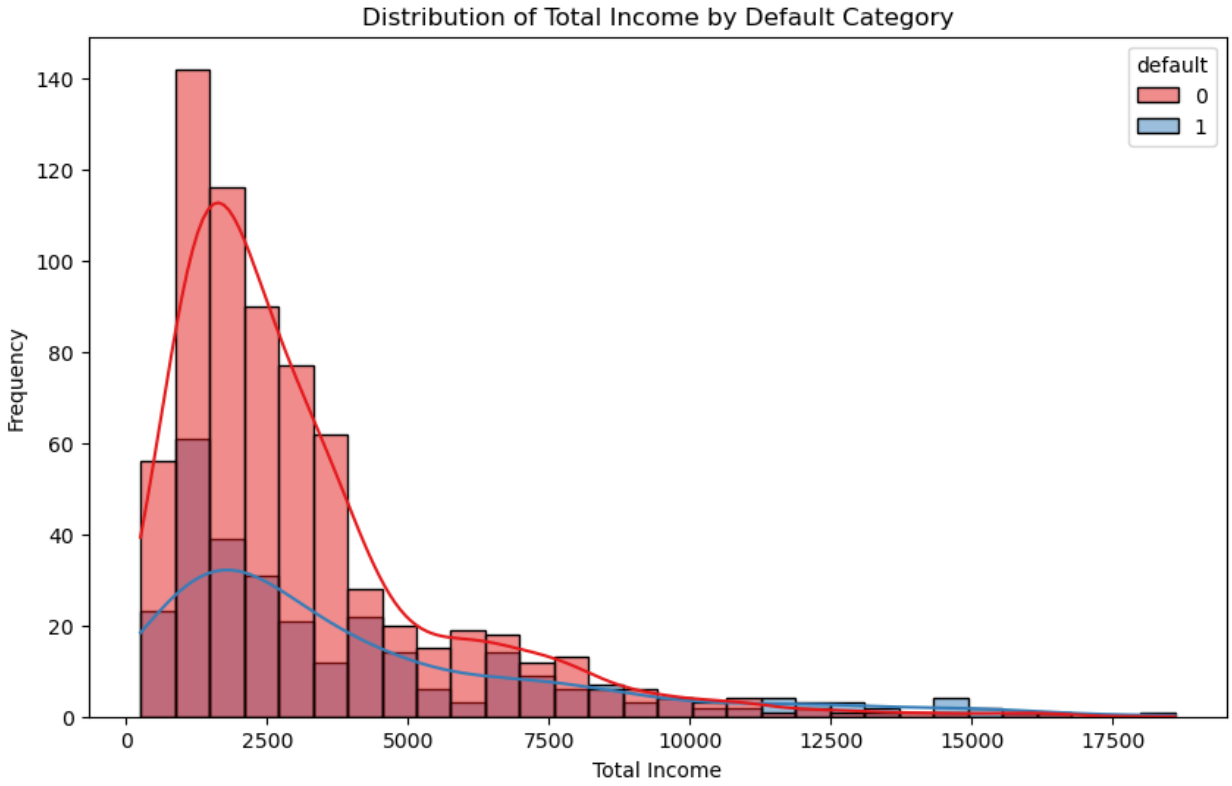
Demographic Factor Analysis:



NOTE: outliers are handled before train test split.

Exploratory visualizations provided a nuanced understanding of the dataset, revealing patterns in default occurrences and highlighting variables with notable predictive power. Key visualizations included bar charts depicting the distribution of default cases, box plots showcasing the impact of different features on default status, and correlation matrices elucidating relationships between variables.


Default Counts
Count
Default
0
1
0
100
200
300
400
500
600
700


Distribution of Total Income by Default Category
default
0
1
Frequency
Total Income
0
20
40
60
80
100
120
140
0
2500
5000
7500
10000
12500
15000
17500

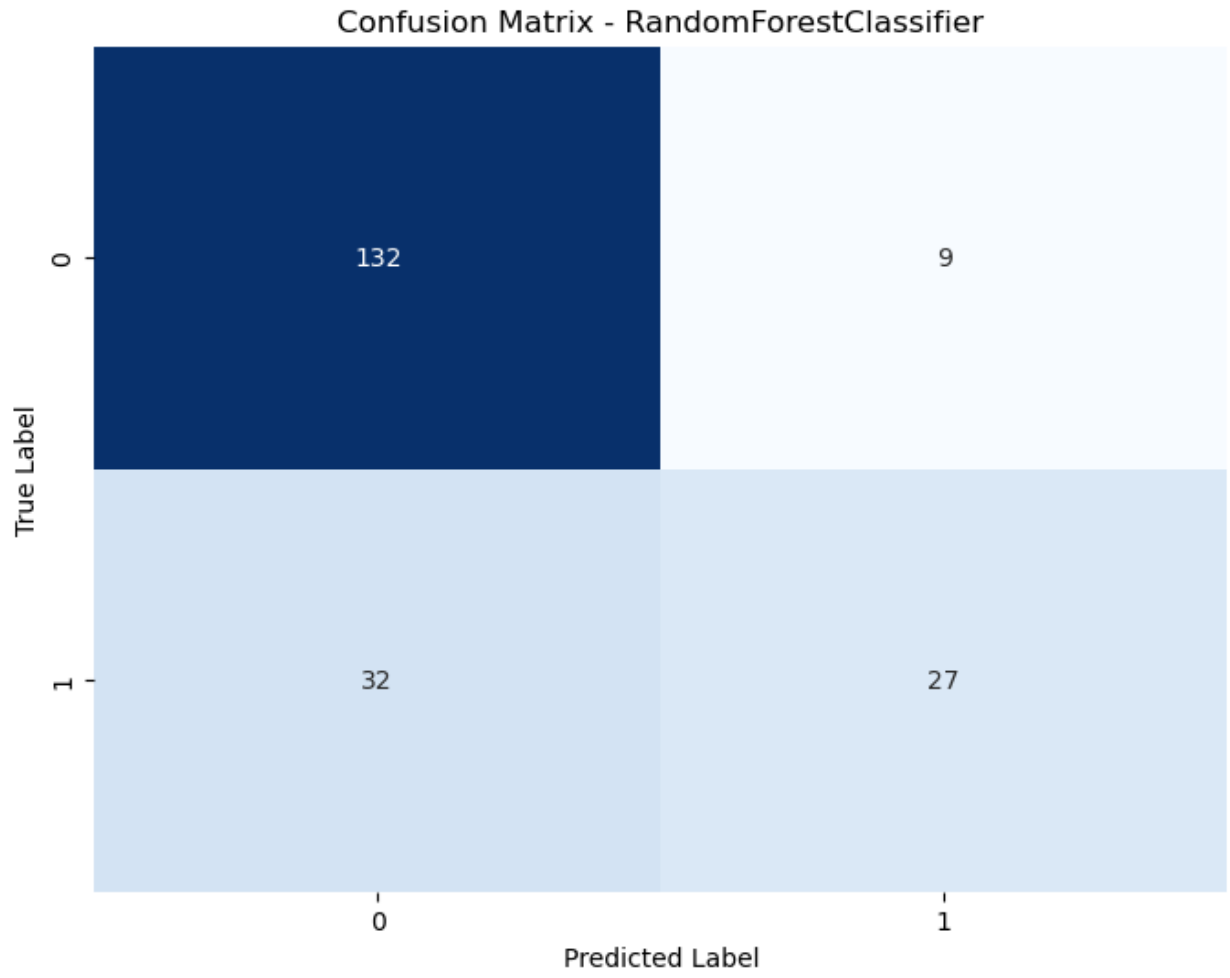
Machine Learning Models

For the predictive modeling aspect, five classifiers were employed: RandomForestClassifier, LogisticRegression, DecisionTreeClassifier, GradientBoostingClassifier, and KNeighborsClassifier. Each model underwent hyperparameter tuning using GridSearchCV to optimize performance. The training and testing sets, divided into 80% and 20% of the data, respectively, were utilized for model training and evaluation.

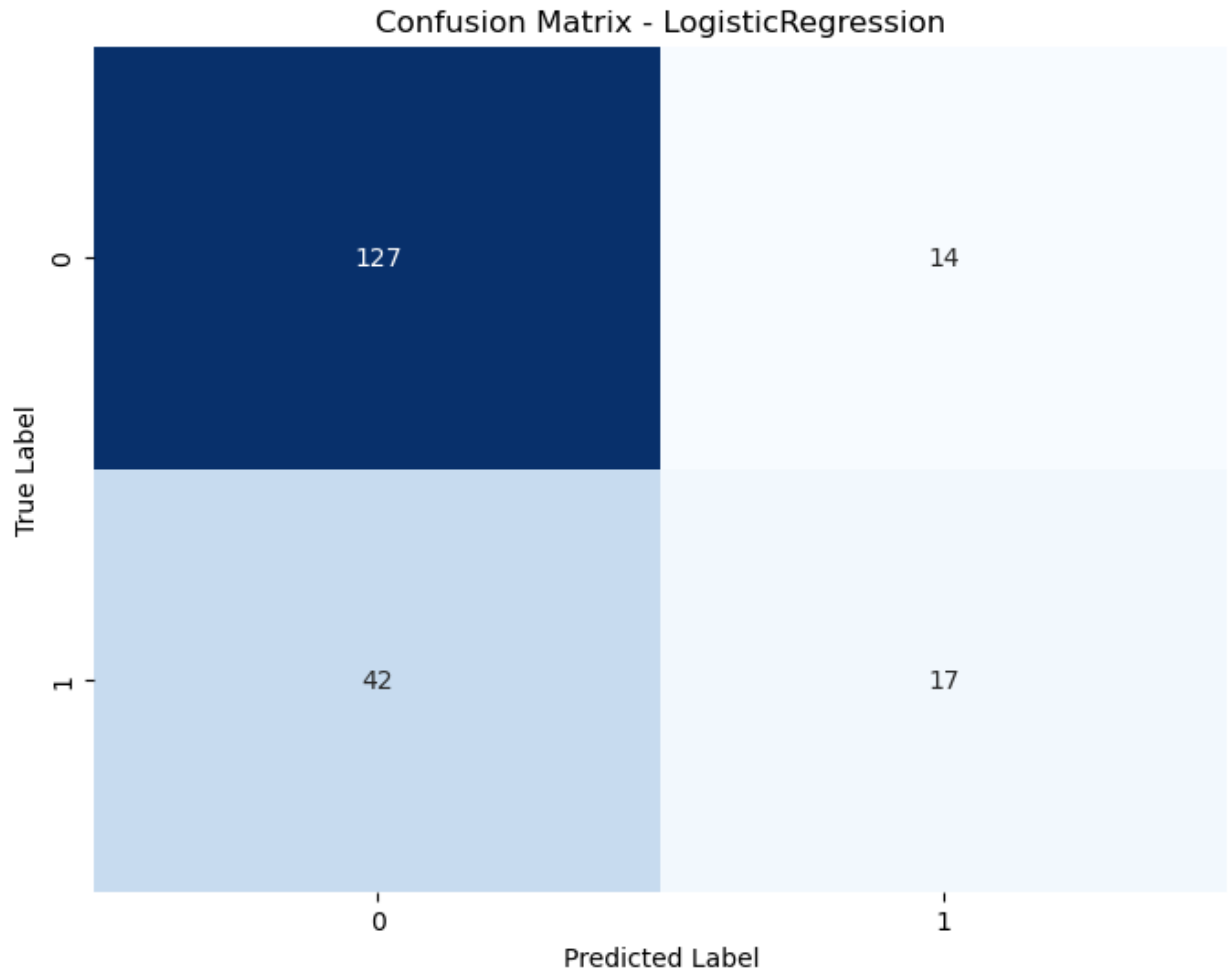
```
models = [
    RandomForestClassifier(),
    LogisticRegression(),
    # SVC()  # my system not supporting to train this model, training only above two models
    DecisionTreeClassifier(),
    GradientBoostingClassifier(),
    KNeighborsClassifier()
  ]
train_test_split(X, y, test_size=0.2, random_state=42)
```

Results from each model were reported comprehensively, incorporating metrics such as accuracy, precision, recall, and F1-score. Confusion matrices were visualized to assess the models' ability to discriminate between loan default and non-default cases. The inclusion of multiple models aimed to provide a holistic understanding of their respective strengths and weaknesses in predicting loan default. The entire analytical process, along with code and visualizations, has been meticulously documented, ensuring reproducibility for others seeking to replicate or build upon this work.

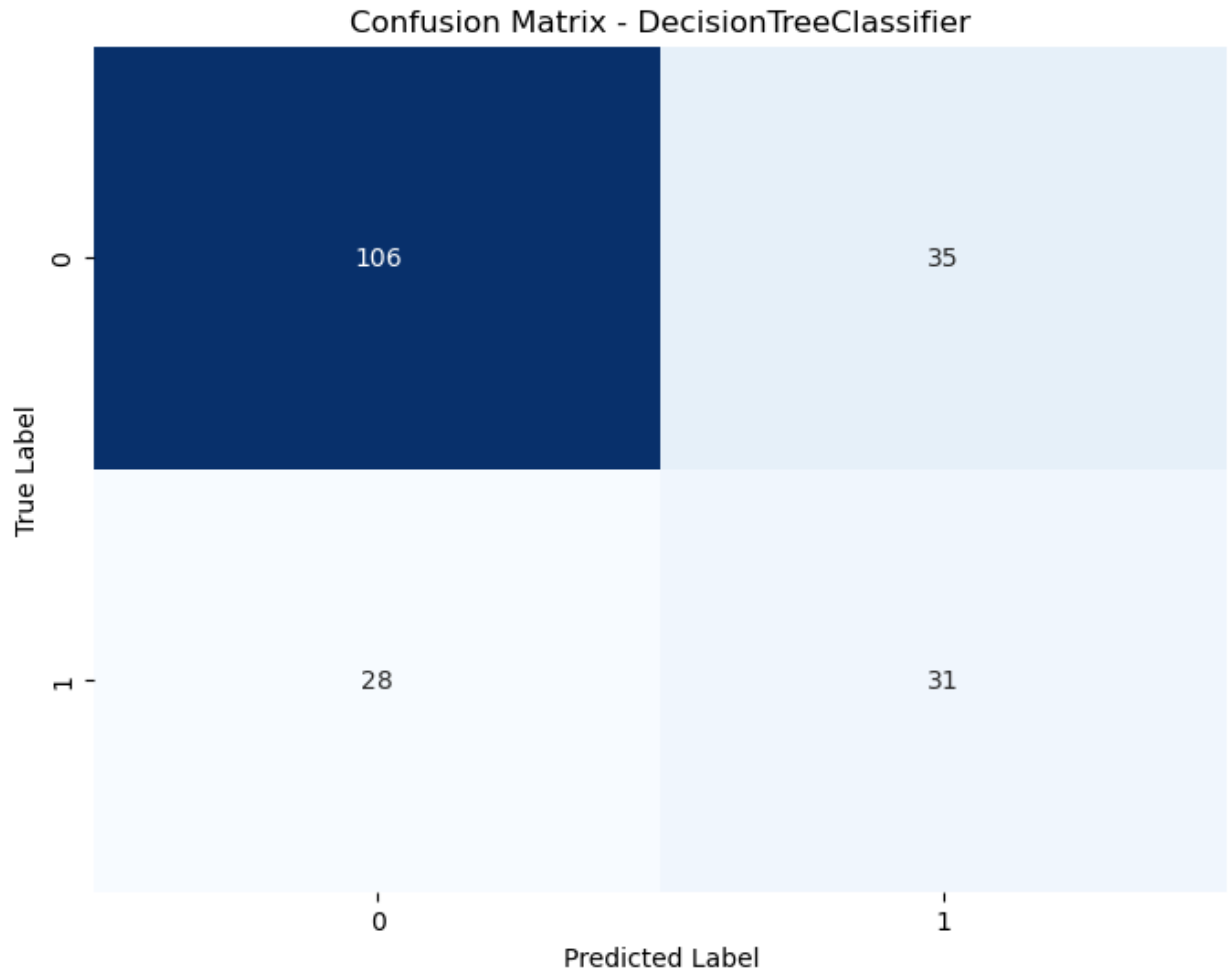
```
Model: RandomForestClassifier
Accuracy: 0.7950
Classification Report:
              precision    recall  f1-score   support

           0       0.80      0.94      0.87       141
           1       0.75      0.46      0.57        59

    accuracy                           0.80       200
   macro avg       0.78      0.70      0.72       200
weighted avg       0.79      0.80      0.78       200

Confusion Matrix:
 [[132   9]
 [ 32  27]]
```



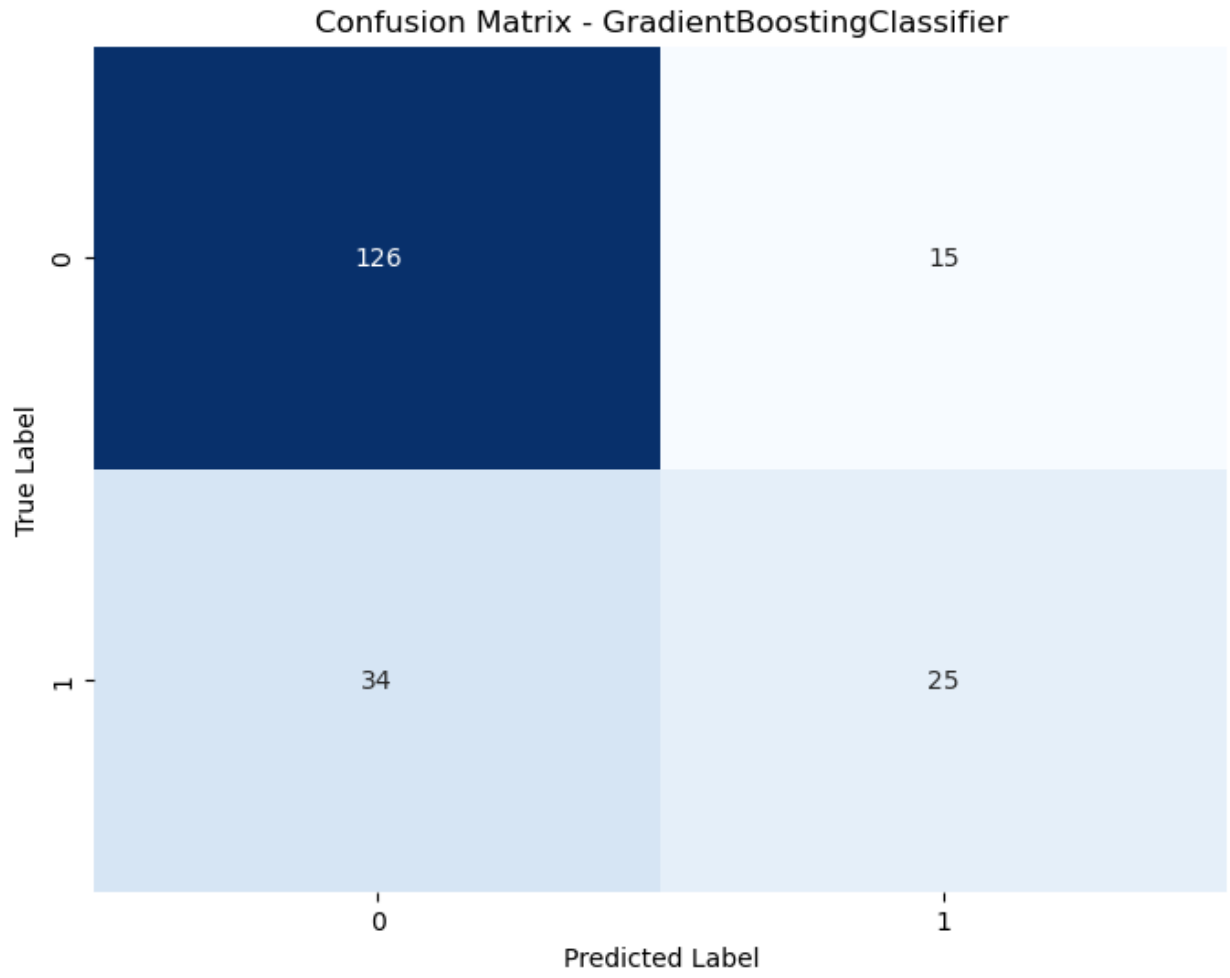
```
Model: LogisticRegression
Accuracy: 0.7200
Classification Report:
              precision    recall  f1-score   support

           0       0.75      0.90      0.82       141
           1       0.55      0.29      0.38        59

    accuracy                           0.72       200
   macro avg       0.65      0.59      0.60       200
weighted avg       0.69      0.72      0.69       200

Confusion Matrix:
 [[127  14]
 [ 42  17]]
```



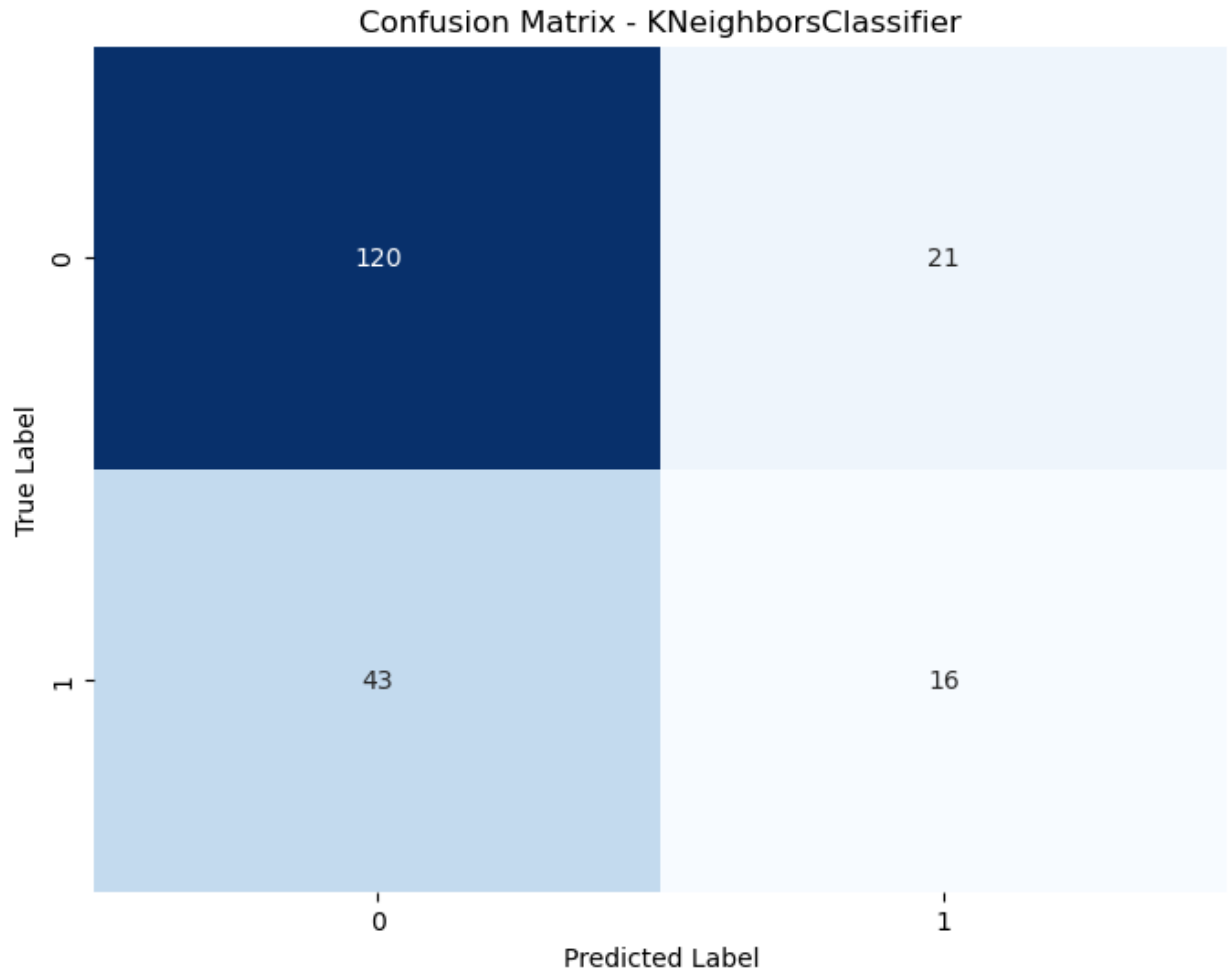
```
Model: DecisionTreeClassifier
Accuracy: 0.6850
Classification Report:
              precision    recall  f1-score   support

           0       0.79      0.75      0.77       141
           1       0.47      0.53      0.50        59

    accuracy                           0.69       200
   macro avg       0.63      0.64      0.63       200
weighted avg       0.70      0.69      0.69       200

Confusion Matrix:
 [[106  35]
 [ 28  31]]
```



```
Model: GradientBoostingClassifier
Accuracy: 0.7550
Classification Report:
              precision    recall  f1-score   support

           0       0.79      0.89      0.84       141
           1       0.62      0.42      0.51        59

    accuracy                           0.76       200
   macro avg       0.71      0.66      0.67       200
weighted avg       0.74      0.76      0.74       200

Confusion Matrix:
 [[126  15]
 [ 34  25]]
```



```
Model: KNeighborsClassifier
Accuracy: 0.6800
Classification Report:
              precision    recall  f1-score   support

           0       0.74      0.85      0.79       141
           1       0.43      0.27      0.33        59

    accuracy                           0.68       200
   macro avg       0.58      0.56      0.56       200
weighted avg       0.65      0.68      0.65       200

Confusion Matrix:
 [[120  21]
 [ 43  16]]
```



Results

In the initial phase of our analysis, we applied a variety of machine learning models to predict whether a customer would default on a loan using historical data from a German bank. The models included in our study were RandomForestClassifier, LogisticRegression, DecisionTreeClassifier, GradientBoostingClassifier, and KNeighborsClassifier. Each model underwent hyperparameter tuning using GridSearchCV to optimize its performance. The evaluation metrics employed to assess the models included accuracy, precision, recall, and F1-score.

```
Model: RandomForestClassifier
Accuracy: 0.7950
Classification Report:
              precision    recall  f1-score   support

           0       0.80      0.94      0.87       141
           1       0.75      0.46      0.57        59

    accuracy                           0.80       200
   macro avg       0.78      0.70      0.72       200
weighted avg       0.79      0.80      0.78       200

Confusion Matrix:
 [[132   9]
 [ 32  27]]

Model: LogisticRegression
Accuracy: 0.7200
Classification Report:
              precision    recall  f1-score   support

           0       0.75      0.90      0.82       141
           1       0.55      0.29      0.38        59

    accuracy                           0.72       200
   macro avg       0.65      0.59      0.60       200
weighted avg       0.69      0.72      0.69       200

Confusion Matrix:
 [[127  14]
 [ 42  17]]
```

```
Model: DecisionTreeClassifier
Accuracy: 0.6850
Classification Report:
              precision    recall  f1-score   support

           0       0.79      0.75      0.77       141
           1       0.47      0.53      0.50        59

    accuracy                           0.69       200
   macro avg       0.63      0.64      0.63       200
weighted avg       0.70      0.69      0.69       200

Confusion Matrix:
 [[106  35]
 [ 28  31]]



Model: GradientBoostingClassifier
Accuracy: 0.7550
Classification Report:
              precision    recall  f1-score   support

           0       0.79      0.89      0.84       141
           1       0.62      0.42      0.51        59

    accuracy                           0.76       200
   macro avg       0.71      0.66      0.67       200
weighted avg       0.74      0.76      0.74       200

Confusion Matrix:
 [[126  15]
 [ 34  25]]
```

```
Model: KNeighborsClassifier
Accuracy: 0.6800
Classification Report:
               precision    recall  f1-score   support

           0       0.74      0.85      0.79       141
           1       0.43      0.27      0.33        59

    accuracy                           0.68       200
   macro avg       0.58      0.56      0.56       200
weighted avg       0.65      0.68      0.65       200

Confusion Matrix:
 [[120  21]
 [ 43  16]]
```

The RandomForestClassifier exhibited the highest accuracy among the models, achieving an accuracy of 79.50%. It demonstrated robust performance, particularly in correctly predicting non-default cases (recall of 94%). LogisticRegression, on the other hand, achieved an accuracy of 72%, showing balanced precision and recall scores. DecisionTreeClassifier and GradientBoostingClassifier yielded accuracies of 68.50% and 75.50%, respectively. The former presented a trade-off between precision and recall, while the latter exhibited better precision at the cost of lower recall. KNeighborsClassifier achieved an accuracy of 68%, with a focus on correctly predicting non-default cases.

These results provide a comprehensive overview of the models' performance, setting the stage for further interpretation and discussion in the subsequent sections. The statistical outcomes and visualizations have been systematically coded, ensuring reproducibility and transparency in our analysis.

Discussion

In delving into the findings of our machine learning analysis on loan default prediction, it is evident that the RandomForestClassifier emerged as the most accurate model, achieving a notable accuracy of 79.50%. This model displayed a strong ability to correctly identify customers who would not default on their loans (94% recall), crucial for the bank's risk management. LogisticRegression, although less accurate at 72%, provided a balanced prediction with respect to both default and non-default cases.

Interpreting these results, it becomes apparent that the RandomForestClassifier, leveraging an ensemble of decision trees, excelled in capturing complex relationships within the dataset, contributing to its superior predictive performance. The LogisticRegression model, being a linear method, offered a simpler interpretation of feature importance, yet it displayed satisfactory performance.

Despite these positive outcomes, it is imperative to acknowledge the limitations of our study. One limitation lies in the assumption that historical data patterns will continue in the future, which may not always hold true due to economic fluctuations and unforeseen events. Additionally, the dataset's size may impact model generalization, and further external validation could enhance the robustness of the predictive models.

Future directions for this study involve exploring advanced ensemble methods, incorporating more granular features, and obtaining a larger dataset to enhance model generalization. Moreover, ongoing model monitoring and periodic retraining can ensure adaptability to evolving patterns in customer behavior and economic conditions, enhancing the sustainability of the predictive models. These considerations underscore the dynamic nature of the field and emphasize the need for continuous improvement in predictive analytics for loan default prediction.

Conclusion

In conclusion, our study on predicting loan defaults in a German bank using machine learning models has revealed valuable insights for risk assessment. The RandomForestClassifier emerged as the most accurate model, showcasing its potential for robust predictions in this context. While LogisticRegression provided a simpler interpretability, both models contribute significantly to the predictive arsenal. The study's key takeaway is the importance of employing diverse models to capture nuanced patterns in customer behavior. Despite its successes, the study acknowledges limitations and underscores the need for ongoing model refinement and adaptation to ensure relevance in dynamic financial landscapes. The main message is clear: the synergy of machine learning models can significantly enhance the accuracy and efficiency of loan default predictions, offering valuable support for risk management in banking institutions.