FINAL PROJECT – GERMAN BANK LOAN

INTRODUCTION

The banking industry faces significant challenges with loan defaulters, which can lead to substantial financial losses. This project aims to leverage machine learning to predict whether a customer will default on their loan based on historical data from a German bank. The dataset includes various customer attributes such as employment duration, existing loans count, savings balance, percentage of income, and age. By accurately predicting loan defaults, banks can better manage risk and make more informed lending decisions.

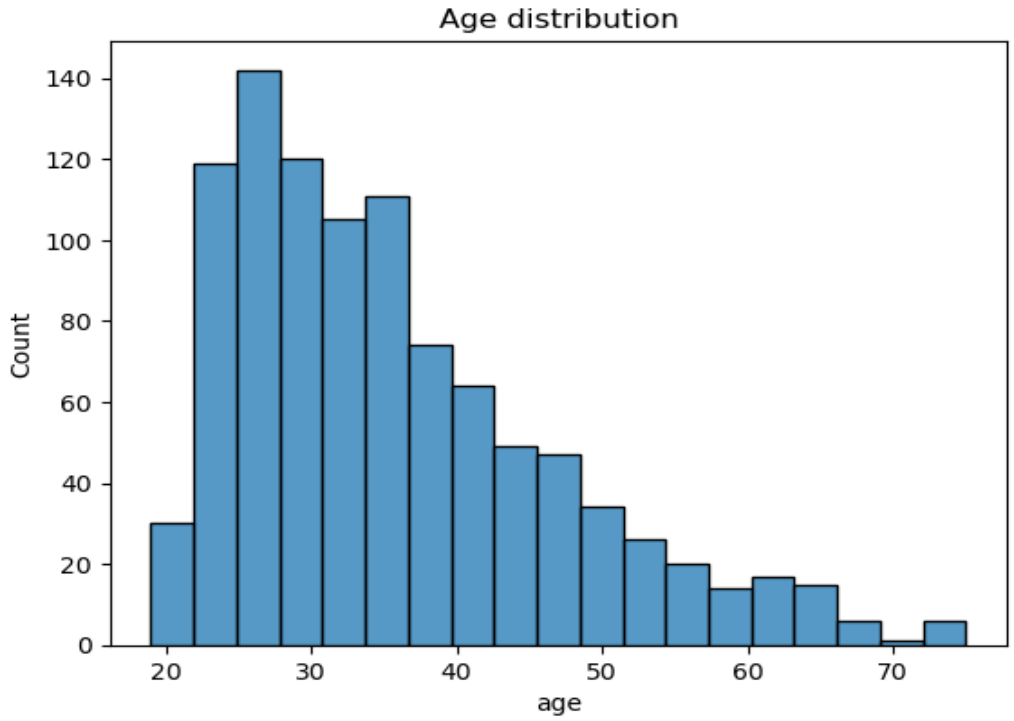
To address this problem, several questions arise: What are the key factors that contribute to loan defaults? Which machine learning models provide the most accurate predictions? How can we interpret the results to improve the bank's decision-making process? This project seeks to explore these questions through data analysis and the application of various machine learning techniques.

METHODS AND MATERIALS

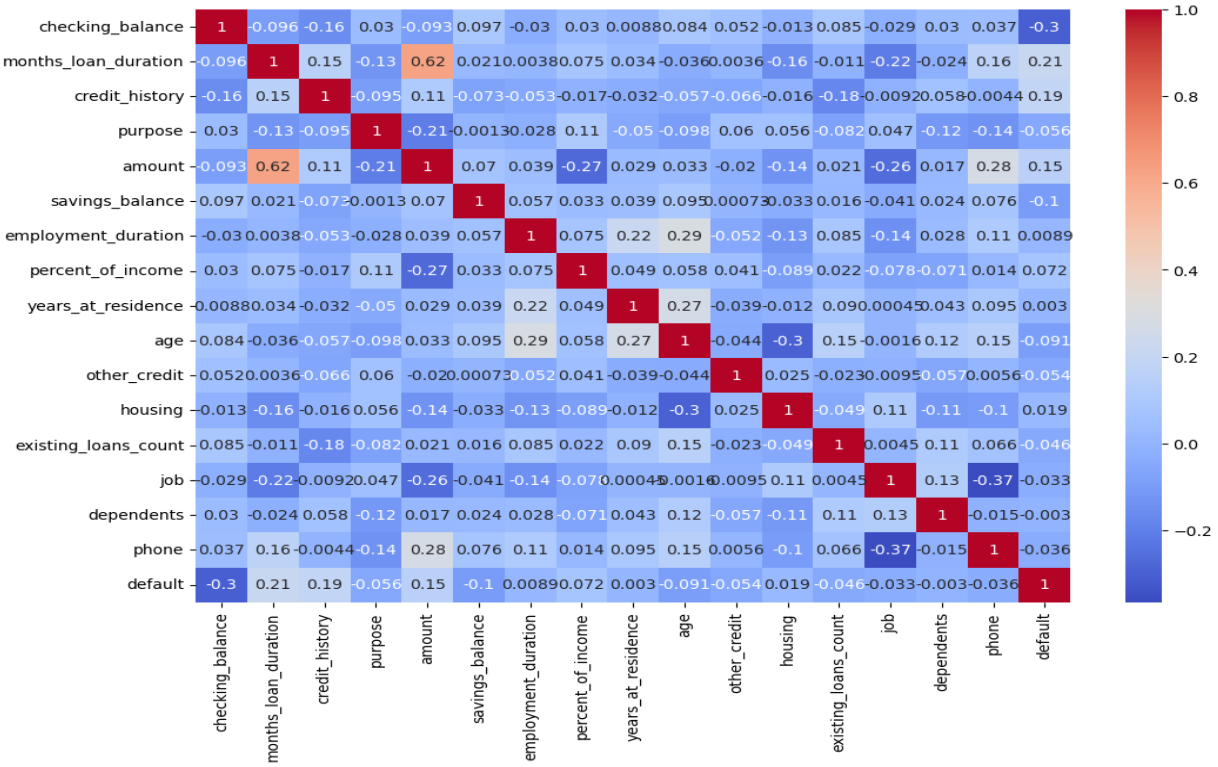
The dataset used in this project, **German_bank.csv**, contains 17 columns and 1000 rows, each representing a customer. Key attributes include checking balance, loan duration, credit history, loan purpose, loan amount, savings balance, employment duration, income percentage, residence duration, age, other credits, housing type, existing loans count, job type, dependents, phone ownership, and default status.

We began with Exploratory Data Analysis (EDA) to understand the dataset's structure and identify any patterns or anomalies. Visualizations such as histograms and heat maps were employed to illustrate distributions and correlations.

Histogram: For example, a histogram of the age distribution to show how customer ages are distributed.



Heatmap: Include the heatmap to show correlations between different features. This helps to identify relationships between variables and determine if there are any multicollinearity issues.



Categorical variables were encoded using LabelEncoder to convert them into numeric form suitable for machine learning algorithms.

Five machine learning models were selected for this study: Logistic Regression, Support Vector Machine (SVM), Random Forest, Gradient Boosting, and Naive Bayes. Each model was trained on 70% of the dataset, with the remaining 30% reserved for testing. The performance of each model was evaluated using confusion matrices and classification reports, focusing on metrics like precision, recall, and f1-score.

RESULTS

```
1.Logistic Regression Results:
[[188  21]
 [ 68  23]]
              precision    recall  f1-score   support

           0       0.73      0.90      0.81       209
           1       0.52      0.25      0.34        91

    accuracy                           0.70       300
   macro avg       0.63      0.58      0.57       300
weighted avg       0.67      0.70      0.67       300


2.SVM Results:
[[209   0]
 [ 86   5]]
              precision    recall  f1-score   support

           0       0.71      1.00      0.83       209
           1       1.00      0.05      0.10        91

    accuracy                           0.71       300
   macro avg       0.85      0.53      0.47       300
weighted avg       0.80      0.71      0.61       300


3.Random Forest Results:
[[188  21]
 [ 51  40]]
              precision    recall  f1-score   support

           0       0.79      0.90      0.84       209
           1       0.66      0.44      0.53        91
```

```
    accuracy                           0.76       300
   macro avg       0.72      0.67      0.68       300
weighted avg       0.75      0.76      0.74       300
```

```
4.Gradient Boosting Results:
[[191  18]
 [ 49  42]]
              precision    recall  f1-score   support

           0       0.80      0.91      0.85       209
           1       0.70      0.46      0.56        91

    accuracy                           0.78       300
   macro avg       0.75      0.69      0.70       300
weighted avg       0.77      0.78      0.76       300
```

```
5.Naive Bayes Results:
[[186  23]
 [ 61  30]]
              precision    recall  f1-score   support

           0       0.75      0.89      0.82       209
           1       0.57      0.33      0.42        91

    accuracy                           0.72       300
   macro avg       0.66      0.61      0.62       300
weighted avg       0.70      0.72      0.69       300
```

DISCUSSION

The results of this study reveal varying performance across the different machine learning models. The Gradient Boosting model achieved the highest accuracy (78%), indicating its effectiveness in capturing complex patterns in the data. This model also demonstrated a balanced precision and recall for both classes. On the other hand, the Support Vector Machine (SVM) showed lower performance, particularly in predicting loan defaults, with a recall of only 0.05 for the default class. This suggests that SVM may not be the best choice for this imbalanced dataset.

The Random Forest model also performed well, showing good precision and recall but slightly lower overall accuracy compared to Gradient Boosting. Naive Bayes, while relatively simple, provided decent results but struggled with predicting defaults effectively. The observed performance gaps highlight the challenges posed by class imbalance and the need for techniques to address this issue, such as SMOTE or other sampling methods.

The limitations of this study include the inherent class imbalance in the dataset, which affected the performance of most models, especially in identifying the minority class (defaults). Future work could explore advanced ensemble methods, feature engineering, or additional data to improve model performance.

CONCLUSION

In conclusion, this project demonstrated the potential of machine learning models to predict loan defaults using historical customer data. The Gradient Boosting model was the most effective, providing the highest accuracy and a balanced performance across classes. Despite the promising results, class imbalance remains a significant challenge, impacting the ability of all models to accurately predict defaults. Future work should focus on addressing these limitations and exploring additional techniques to enhance model performance. Overall, effective prediction of loan defaults can help banks manage risk better and make more informed lending decisions.