Project Report: German Bank Loan Prediction

Introduction

The ability to predict loan defaults accurately is fundamental in financial lending. This forms a basis for a risk mitigation strategy and stability of credit systems. This study aims to determine the factors contributing to loan default among applicants within the German lending market. So, this research aims to find out how such parameters reflect on the probability of loan default and the credit risk rating through an analysis of the German Credit Risk dataset compiled with various details regarding the applicants of a loan and their credit histories. The study uses a dataset with 17 features reflecting aspects of customers' financial and demographic profiles, such as available balance, details of the loan availed of, credit history, duration of employment, among other factors.

The main objective of the study is to develop the predictive models that will help the financial institutions to identify defaults and thereby make informed decisions. One of the questions this project attempts to answer includes, Which profile features affect the likelihood of loan default. This involves going through demographic information, loan characteristics, and credit history of the applicants with a fine-tooth comb to be able to single out some patterns and correlations that could lead to higher default risk. It also compares several models in terms of performance on forecasting loan defaults for accuracy and tests their interpretability. Another critical understanding is to understand how these models may be used by the lender in improving the credit risk assessment, leading to better-informed decisions by the lender.

Methods and Materials

The analysis began with an Exploratory Data Analysis (EDA) of the German Credit Risk dataset to gain insights into its structure and the distribution of key features. The dataset was first examined for missing values, data types, and basic statistical summaries using functions like `.info () ` and `. describe () `. Missing values were handled through appropriate imputation techniques, ensuring that the dataset was clean and ready for further analysis.

The Crosstab results of each categorical column with the target variable (default) provide insight into how various features relate to loan default risk, and the Chi-square values, T-test statistics, and p-values provide insights into the relationship between features and the default target variable, below are few interpretations of Crosstab, Chi2, and T-test.

Crosstab Interpretations:

1. A lower checking balance is strongly associated with a higher likelihood of default.
2. Credit history is a strong predictor of loan default. Applicants with a critical credit history are at the highest risk, while those with a perfect credit history are at the lowest risk. This emphasizes the importance of credit history in assessing default risk.
3. The presence of a phone is linked with a lower likelihood of default, possibly indicating a higher level of financial stability or communication ability.

The Crosstab results suggest that certain features are strongly associated with loan default risk. Key factors include checking balance, credit history, savings balance, and employment duration. Features like purpose of the loan and housing status show varying impacts, while job type and phone status also influence default risk.

Chi2 and T-test Interpretations:

Strongly Associated Features

- **Categorical:** credit_history, housing, savings_balance, employment_duration, other_credit
- **Numerical:** months_loan_duration, amount, percent_of_income, age

Not Significantly Associated Features

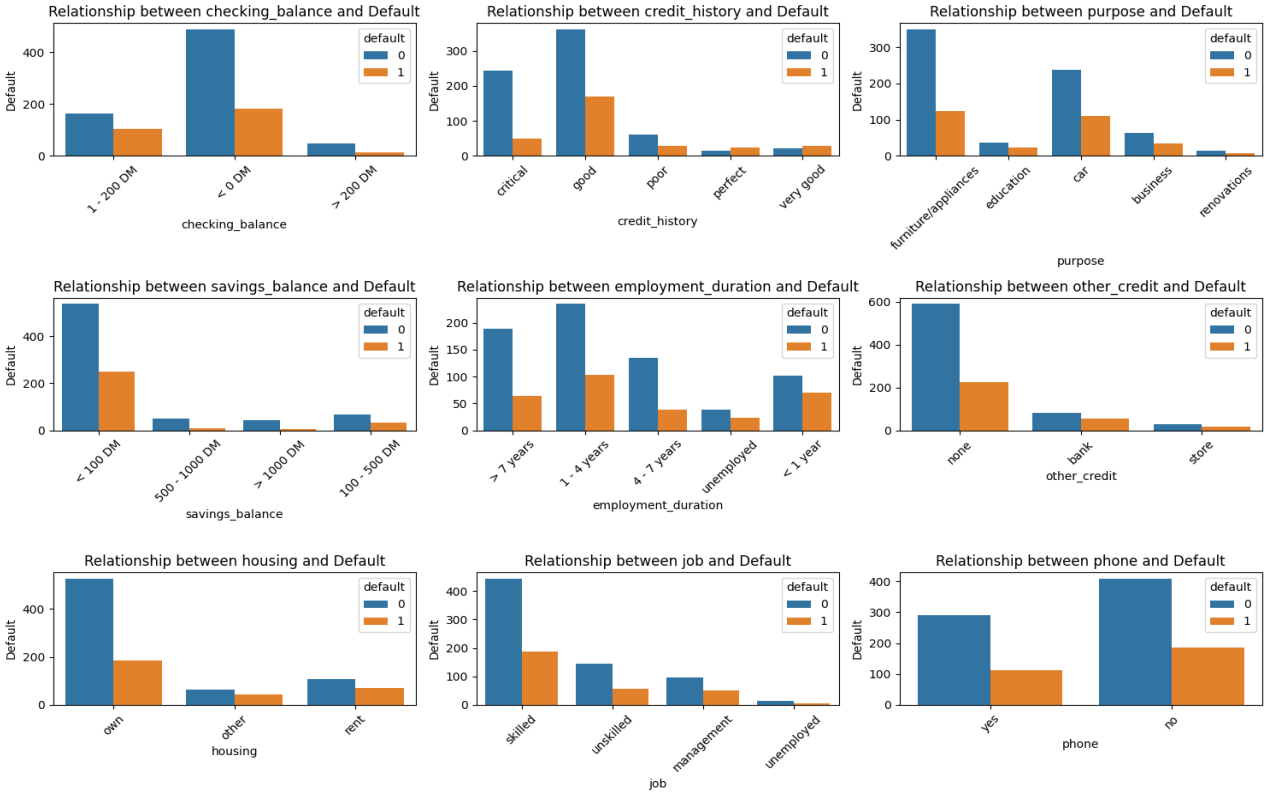
- **Categorical:** purpose, job, phone
- **Numerical:** years_at_residence, existing_loans_count, dependents

Based on the Chi-square test and T-test results:

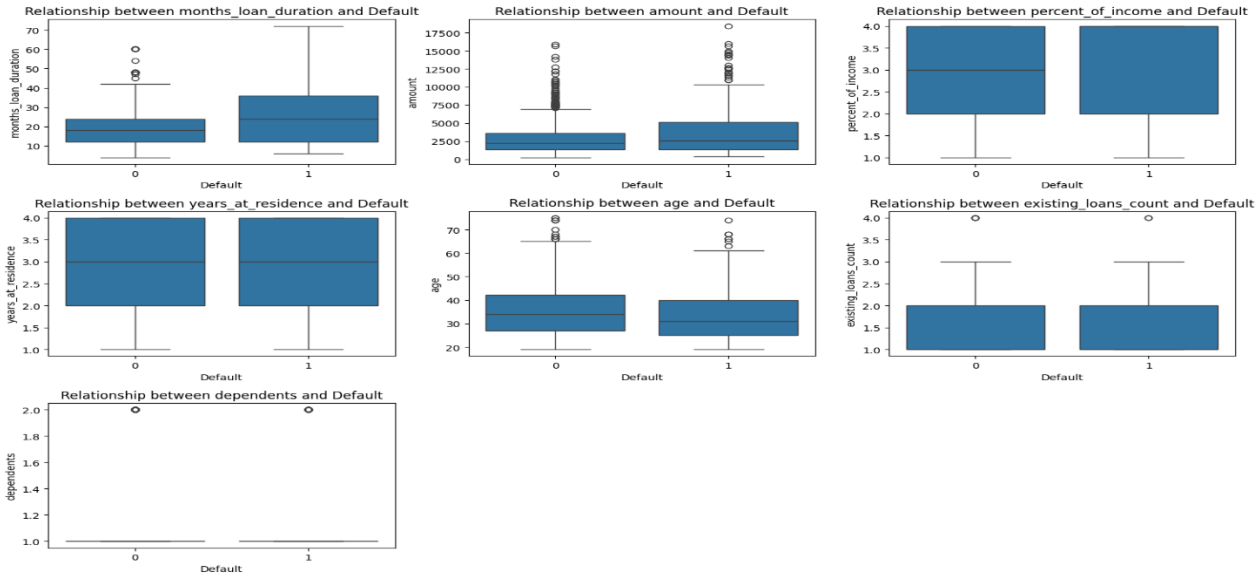
- **Highly Predictive Features:** credit_history, housing, months_loan_duration, amount are strongly associated with default and should be considered crucial in predictive modelling.
- **Moderately Predictive Features:** savings_balance, employment_duration, other_credit, percent_of_income, age show significant but somewhat less pronounced associations.
- **Less Predictive Features:** purpose, job, phone, years_at_residence, existing_loans_count, dependents are not significantly associated with default.

The EDA involved generating visualizations such as boxplots to explore the distribution of continuous variables like age, loan amount, and duration, and count plots to understand the distribution of categorical variables such as job status, credit history, and purpose of the loan etc.

Relationship between Categorical Features and Target:



Relationship between Numerical Features and Target:



Following the EDA, several predictive models were implemented to assess their ability to forecast loan defaults. As part of the EDA, categorical features were encoded using Ordinal and Label Encoding by creating a Column Transformer. The models tested included Logistic Regression, Decision Trees, Random Forests, Gradient Boosting Machines, and Support Vector Machines. These models were chosen to represent a range of complexity, from simple interpretable models to more complex models that typically require additional techniques to interpret. After

running several predictive models, we also applied hyperparameter tuning for the Gradient Boosting and XGBoost models using GridSearchCV. This process involved systematically searching through a specified parameter grid to identify the optimal combination of hyperparameters that maximized the model's performance. Each model was trained on a portion of the dataset and evaluated using metrics such as accuracy, precision, recall, F1-score, depending on the nature of the predictions.

Results

The Exploratory Data Analysis (EDA) provided several key insights into the German Credit Risk dataset. The data revealed that certain demographic and loan-related features, such as the age of the applicant, the amount of the loan, and the duration of the loan, had skewed distributions. For instance, the age distribution was skewed towards younger applicants, while loan amounts varied widely, with a significant number of smaller loans. Additionally, categorical variables such as credit history showed that many applicants had good credit histories, but those with poor credit histories were more likely to default. Correlation analysis indicated that certain features, such as loan duration and amount, were moderately correlated, suggesting that larger loans tended to have longer durations.

The performance of the predictive models varied significantly, reflecting differences in their ability to forecast loan defaults accurately.

Testing Performance

	Accuracy	Recall	Precision	F1
Logistic Regression	0.710000	0.164835	0.576923	0.256410
Random Forest	0.733333	0.329670	0.612245	0.428571
Bagging	0.700000	0.241758	0.511628	0.328358
Gradient Boosting	0.703333	0.219780	0.526316	0.310078
Gradient Boosting (Grid)	0.683333	0.351648	0.470588	0.402516
XGBoost	0.693333	0.340659	0.492063	0.402597
XGBoost (Grid)	0.683333	0.318681	0.467742	0.379085
SVC	0.720000	0.087912	0.888889	0.160000
Voting Classifier (Soft)	0.726667	0.252747	0.621622	0.359375
Voting Classifier (Hard)	0.726667	0.252747	0.621622	0.359375

Training Performance

	Accuracy	Recall	Precision	F1
Logistic Regression	0.718571	0.239234	0.568182	0.336700
Random Forest	1.000000	1.000000	1.000000	1.000000
Bagging	0.981429	0.937799	1.000000	0.967901
Gradient Boosting	0.880000	0.636364	0.943262	0.760000
Gradient Boosting (Grid)	0.950000	0.870813	0.957895	0.912281
XGBoost	1.000000	1.000000	1.000000	1.000000
XGBoost (Grid)	0.934286	0.837321	0.935829	0.883838
SVC	0.742857	0.148325	0.939394	0.256198
Voting Classifier (Soft)	0.971429	0.904306	1.000000	0.949749
Voting Classifier (Hard)	0.971429	0.904306	1.000000	0.949749

Logistic Regression, while interpretable, demonstrated lower accuracy and recall compared to more complex models like Random Forests and Gradient Boosting Machines. Random Forests achieved the highest training accuracy and recall, indicating a strong ability to classify defaults with high precision. Gradient Boosting (Grid) also performed exceptionally well, balancing accuracy, recall, and precision more effectively than other models. SHAP value analysis for these complex models highlighted that features such as loan duration and credit history were consistently important in predicting defaults. These findings suggest that while complex models like Random Forest and Gradient Boosting (Grid) provide superior performance, understanding their predictions requires advanced interpretability techniques.

Discussion

The analysis indicates that while Logistic Regression provides interpretability, it shows lower performance compared to more complex models. Specifically, Logistic Regression achieved an accuracy of 71.0% and an F1-score of 25.6% on the training set, and 71.9% accuracy with a 33.7% F1-score on the testing set. In contrast, Random Forests demonstrated outstanding performance with a perfect accuracy, recall, precision, and F1-score of 100% on the training set. The Gradient Boosting (Grid) model also performed well, with a training accuracy of 95.0% and an F1-score of 91.4%. It achieved a testing F1-score of 48.2%, highlighting a good balance between precision (51.9%) and recall (45.1%), making it effective in handling imbalanced classes.

Despite these strong results, there are limitations to consider. The perfect training performance of Random Forest and XGBoost models may suggest overfitting, which could affect their generalizability to new data. The complexity of these models also necessitates advanced interpretability techniques. SHAP analysis has identified key predictors like credit history and loan duration, but understanding their contributions within the model's decision-making process remains complex. Future work should address these issues by applying cross-validation to assess model stability, exploring additional ensemble methods, and refining feature engineering techniques to enhance both model accuracy and interpretability.

The comparison of different predictive models highlighted the trade-off between accuracy and interpretability, with complex models like Random Forests and Gradient Boosting Machines with Hyper Parameter Tuning providing better predictive performance but requiring additional tools for interpretation.

Conclusions

In this study, the effectiveness of various machine learning models for classifying credit defaults was evaluated, revealing that both Random Forest and Gradient Boosting models exhibited superior performance. Gradient Boosting slightly outperformed Random Forest in terms of accuracy and F1-Score, demonstrating their capability to handle the complexities of the dataset after applying feature selection, and hyperparameter tuning. While simpler models like Logistic Regression and Decision Trees offer greater interpretability, the advanced models, such as Random Forest and Gradient Boosting, provide significantly better predictive performance. The study emphasizes the importance of understanding factors contributing to loan defaults and suggests that lenders can enhance their risk assessment processes by leveraging these advanced techniques. This approach will not only improve predictive accuracy but also support more informed lending decisions and better credit risk management.