Name: Pooja Pandit

Name of the Professor: Dr. Cristian Roman Palacios

Subject name: Introduction to Machine Learning

21st August 2024

Final Report

Predicting Loan Default Using Different Machine Learning Models

Introduction

The risk of loan defaults poses a significant challenge to financial institutions, as it directly affects their profitability and stability. Accurately predicting whether a customer will default on a loan allows banks to manage their risk more effectively and make more informed lending decisions. This project explores the use of various machine learning models to predict loan default based on historical customer data from a German bank.

The dataset includes information on customer demographics, financial status, and loan history. The primary objective of this study is to identify the best-performing model that can be used to predict loan defaults accurately.

The specific questions addressed in this study are:

1. Which machine learning model provides the most accurate predictions of loan default?
2. How do different features (such as credit history, loan amount, and employment duration) impact the likelihood of default?

3. What are the trade-offs between different performance metrics (e.g., accuracy, precision, recall) when selecting the best model?

Methods and Materials

Data Collection and Preprocessing

The dataset used in this study contains 1,000 rows and 17 columns, each representing different customer attributes. The target variable **'default'**, indicates whether a customer has defaulted on their loan. The dataset includes a mix of numerical and categorical variables. The preprocessing steps involved:

- **Handling Missing Data And Duplicates**: Checked for missing values. There were no missing values. Additionally checked if there exists any duplicate rows.
- **Encoding Categorical Variables**: Categorical variables were encoded using one-hot encoding, and ordinal variables were encoded according to their hierarchical nature.
- **Standardization**: Numerical features were standardized to ensure that they were on a similar scale, which is particularly important for models sensitive to feature scaling like SVM and KNN.

Feature Engineering

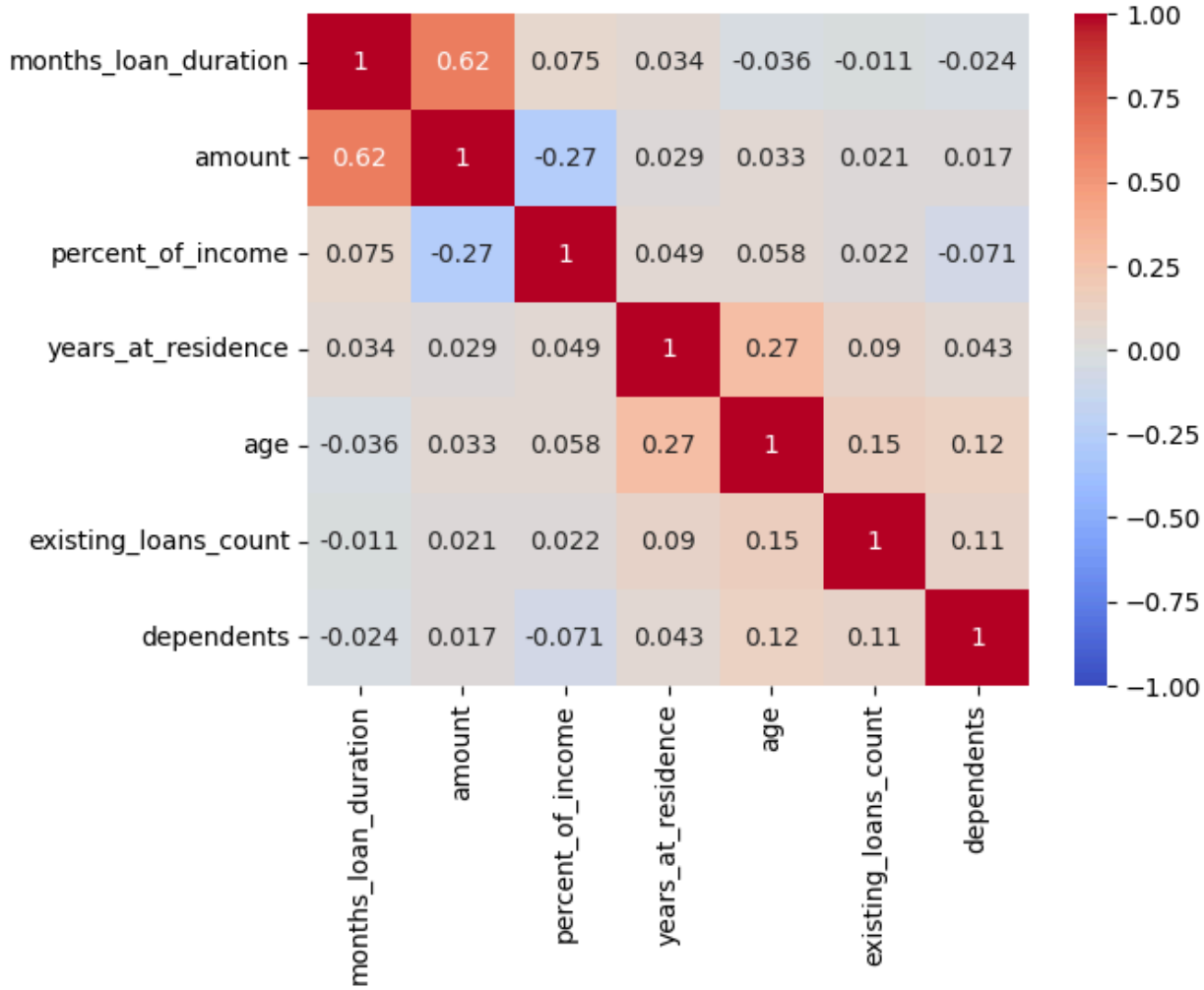
To enhance the predictive power of the models, features such as credit_history, employment_duration, and checking_balance were treated as ordinal variables. This treatment

allowed us to capture the inherent order within these features, which could have a significant impact on the model's ability to predict defaults.

Exploratory Data Analysis (EDA) and Data Visualization

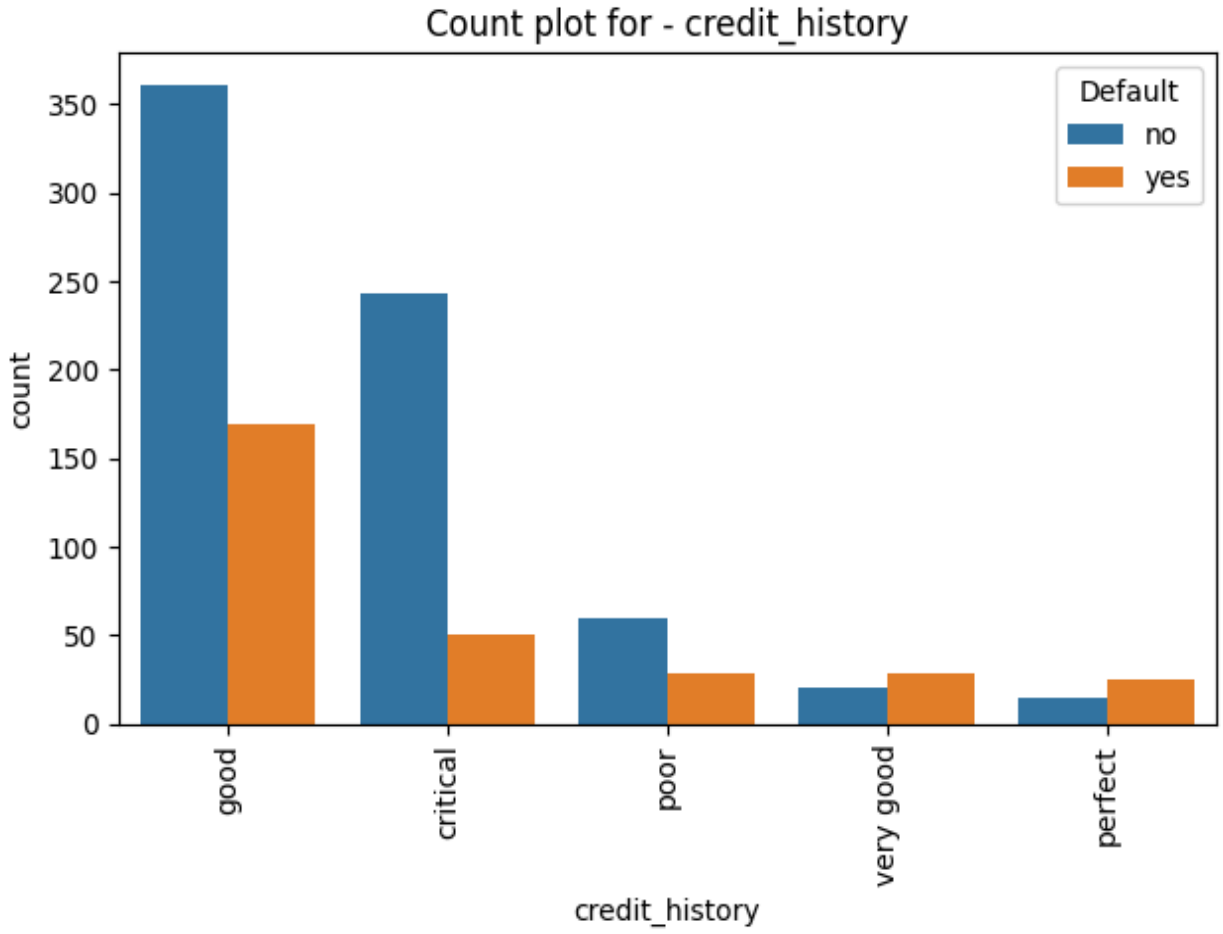
Appropriate visualization plots are created to understand and make the right observations. EDA is performed to have a better understanding of the data distribution and also the relationship of the features of the dataset.

Heatmap highlighting the correlation between different numeric attributes:



Observations for numeric attributes:

The numeric columns are not highly correlated to each other and the dataset considering only numeric columns is non-linear in nature.



Observation for 'credit_history' attribute:

An interesting yet contradictory to normal understanding, I see that people with very good credit history seem to be the most in numbers as loan defaulters. Additionally, people with perfect credit history also are more in numbers when considered as loan defaulters.

There are many plots created to perform EDA and data visualization to better understand the dataset in the python code file provided (.ipynb).

Model Selection and Training

A variety of machine learning models were selected for this study, including:

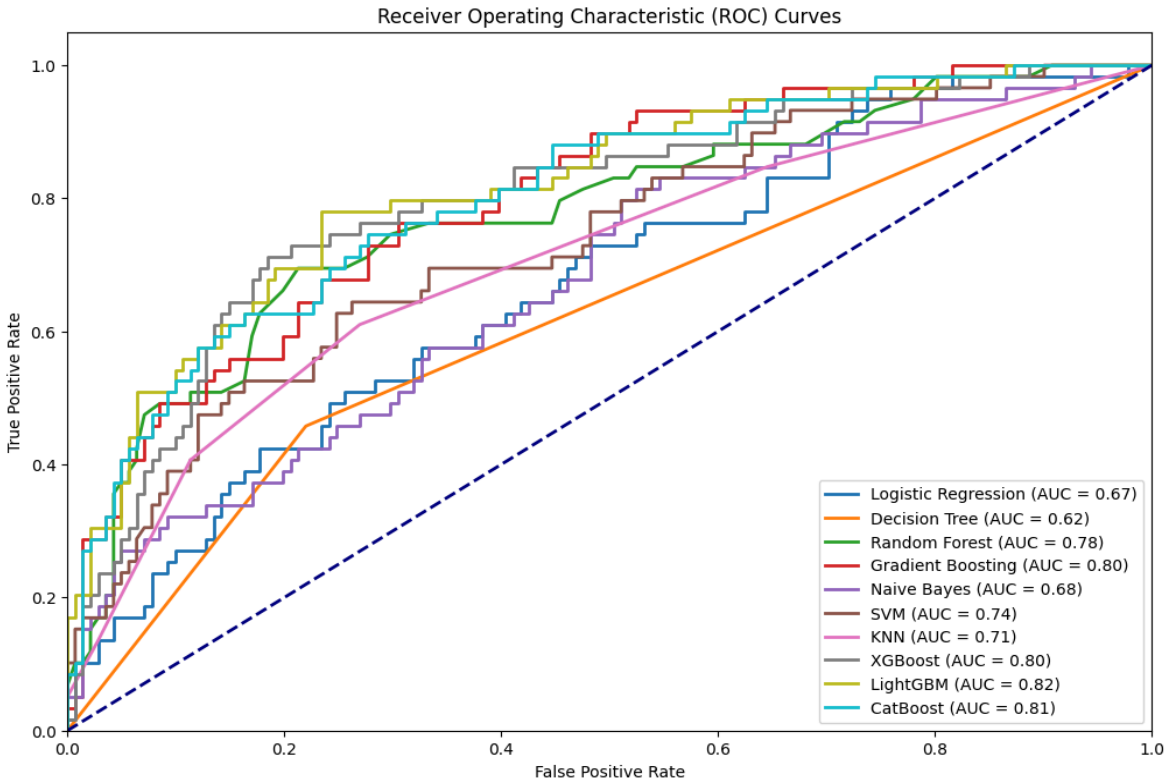
1. **Linear Regression**
2. **Logistic Regression**
3. **Decision Tree**
4. **Random Forest**
5. **Gradient Boosting**
6. **Naive Bayes**
7. **Support Vector Machine (SVM)**
8. **K-Nearest Neighbors (KNN)**
9. **XGBoost**
10. **LightGBM**
11. **CatBoost**

Before feeding the training and test data to each model, each attribute value was converted into numeric data type and then standardized. Later, each model was trained on the preprocessed dataset.

Performance Metrics

The models were evaluated using the following performance metrics:

- **Accuracy**: The proportion of correct predictions out of all predictions made.
- **Precision**: The proportion of true positive predictions out of all positive predictions made by the model.
- **Recall**: The proportion of true positive predictions out of all actual positives in the dataset.
- **F1-Score**: The harmonic mean of precision and recall, providing a balance between the two.
- **ROC-AUC**: The area under the Receiver Operating Characteristic curve, which measures the model's ability to distinguish between the classes.



Results

The performance of each model is summarized in the table below:

	Model	Accuracy	Precision (1)	Recall (1)	F1-Score (1)	ROC-AUC
0	Linear Regression	0.715	0.54	0.25	0.34	0.67
1	Logistic Regression	0.715	0.53	0.27	0.36	0.67
2	Decision Tree	0.685	0.47	0.46	0.46	0.62
3	Random Forest	0.780	0.74	0.39	0.51	0.78
4	Gradient Boosting	0.785	0.71	0.46	0.56	0.80
5	Naive Bayes	0.665	0.44	0.46	0.45	0.68
6	SVM	0.735	0.64	0.24	0.35	0.74
7	KNN	0.745	0.60	0.41	0.48	0.71
8	XGBoost	0.770	0.64	0.49	0.56	0.80
9	LightGBM	0.805	0.76	0.49	0.60	0.82
10	CatBoost	0.785	0.72	0.44	0.55	0.81

Key Findings

- **LightGBM** achieved the highest accuracy (0.805) and ROC-AUC score (0.82), making it the best-performing model overall.
- **Gradient Boosting** and **CatBoost** also performed well, with strong F1-Scores and ROC-AUC values, indicating their potential as reliable models for this task.
- **Random Forest** provided a good balance between precision and recall, with a relatively high accuracy (0.780) and a reasonable ROC-AUC score (0.78).

- **Naive Bayes** and **SVM** had lower recall rates, suggesting they may not be the best choices for identifying defaulters in this dataset.

Discussion

The results of this study indicate that ensemble models, particularly **LightGBM**, are highly effective at predicting loan defaults. The high ROC-AUC score of LightGBM suggests that it has a strong ability to distinguish between defaulters and non-defaulters. However, while models like **Logistic Regression** and **Decision Tree** are simpler, they showed lower performance in this context, highlighting the importance of using more sophisticated models for complex tasks like default prediction.

One of the limitations of this study is the size and scope of the dataset. With only 1,000 rows, the models might not fully generalize to larger, more diverse populations. Future work could involve expanding the dataset, performing effective hyperparameter tuning, incorporating additional features, and exploring more advanced techniques such as deep learning.

The imbalance in the dataset (i.e., fewer defaulters compared to non-defaulters) may have affected the performance metrics, particularly recall. Addressing this imbalance through techniques like SMOTE (Synthetic Minority Over-sampling Technique) could improve the model's ability to identify defaulters.

Conclusion

This study demonstrates the effectiveness of machine learning models in predicting loan defaults, with **LightGBM** emerging as the best-performing model. The ability to accurately predict defaults can provide significant value to banks by helping them mitigate risk. Future research should focus on improving model generalizability and addressing the limitations discussed.