|  |
| --- |
| Machine learning approach to predict bank default |

The report represents a machine learning project aimed at predicting loan defaults using historical data from a German bank. The study involved data exploration, preprocessing data, performing exploratory data analysis, and apply various machine learning models to predict default probabilities. The model’s performance was evaluated using accuracy score and ROC\_AUC curve. The Ada Boost Classifier and Random Forest Classifier was found to be the most effective model with accuracy 0.74 on the test set and ROC 0.79 and 0.78 respectively. The finding suggests that machine learning can be evaluable tool for financial situation in assessing credit risks.

## Introduction

The financial industry has long faced challenges in assessing the creditworthiness of loan applicants. With the advents of machine learning, banks now have an opportunity to improve their prediction of loan defaults, thereby reducing financial risks. This project focus on building and comparing machine learning models to predict whether customers of German bank will default on their loans.

Primary objective of this study is to identify the most effective machine learning model for predicting loan defaults based on historical customer data. By achieving accurate predictions, the bank can make more landing decisions and minimize the incidence of loan defaults.

This report details the methodology used in building the predictive models, the results obtained, and the implications of the finding for credit risk assessment in the bank sector.

## **Methods and Materials**

The dataset used on this study “credit.csv”, contains 1000 entries with 16 features and 1 target feature (“default”), related to customer’s demographic and account information, as well as their loan repayments status. The data was preprocessed to handle missing values, encode categorical variables, and normalize the features. Detail of data description were showed in Appendix A table1 and table2.

Exploratory data analysis revealed key insight into the relationship between various features and target variable (loan default). Visualization such as histogram and correlation matrix were used to understand the data distribution and features importance. Data visualization can be found at Appendix B image1 and image2.

Several machine learning models were evaluated, including Logistic Regression, K-Nearest Neighbors, Support Vector Machines, Linear Discriminant Analysis, Quadratic Discriminant Analysis, ensemble methods like Random Forest, Gradient Boosting, AdaBoost Classifier, XGB Classifier, LGBM Classifier, Bagging Classifier and Voting Classifier method. Hyperparameters tuning was performed by using GridSearch with cross-validation to optimize each model performance.

The models were trained on 70% of data and tested on remaining 30%. Performance score like accuracy and ROC\_AUC were used to assess each model’s effectiveness in predicting loan default.

## Results

The Ada Boost and Random Forest Classifier achieved the highest accuracy score of 0.74 on the test set with ROC of 0.79 and 0.79, Bagging and XGB Classifier showed highest accuracy score 0.753 however, ROC are low of 0.76 and 0.77. Ensemble methods showed considerable higher accuracy than non-tree-based methods. K-nearest Neighbors showed the lowest accuracy 0.70 with ROC of 0.66.

A comparison of the model’s ROC curves indicated that the ensemble methods outperformed the simpler models in terms of both precision and recall. Detail results, included train and test accuracy, ROC\_AUC score and ROC curve for each model are presented in appendix C.

## Discussion

The superior performance of Ada Boost and Random Forest Classifier suggested that ensemble methods, which combine multiple weak learners to form a strong predictor, are well suited for the problem of loan default prediction. The model’s ability to capture complex non-linear relationships in the data likely contributed to its effectiveness.

One limitation of this study is the data set is relatively small, which may impact the generalizability of the findings. Additionally, the model may benefit from the inclusion of more diverse features like economic indicators or credit history.

Future work could explore the use of deep neural network techniques to explore the enhancement of model accuracy. Further research could also investigate to understand the impact of class imbalance on model performance.

## Conclusion

This study demonstrates the potential of machine learning models to predict the loan default with the high degree of accuracy. Ada Boost and Random Forest Classifier, in particular, proved to be effective tools for credit risk assessment. The finding of this project could assist the financial institutions in making more informed landing decision and ultimately reduce the incidence of loan defaults.

## References

[1] Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning. Springer.

[2] Brownlee, J. (2016). Machine Learning Mastery. Retrieved from <https://machinelearningmastery.com/>

[3] Cristian Román-Palacios. Introduction to machine learning course from University of Arizona.

## Appendix A: Exploratory Data Analysis – Statistical data.

**Table1**: Numerical feature’s statistical data

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Features** | **count** | **mean** | **std** | **min** | **25%** | **50%** | **75%** | **max** |
| months\_loan\_duration | 1000 | 20.90 | 12.06 | 4 | 12 | 18 | 24 | 72 |
| amount | 1000 | 3271.26 | 2822.74 | 250 | 1365.5 | 2319.5 | 3972.25 | 18424 |
| percent\_of\_income | 1000 | 2.97 | 1.12 | 1 | 2 | 3 | 4 | 4 |
| years\_at\_residence | 1000 | 2.85 | 1.10 | 1 | 2 | 3 | 4 | 4 |
| age | 1000 | 35.55 | 11.38 | 19 | 27 | 33 | 42 | 75 |
| existing\_loans\_count | 1000 | 1.41 | 0.58 | 1 | 1 | 1 | 2 | 4 |
| dependents | 1000 | 1.16 | 0.36 | 1 | 1 | 1 | 1 | 2 |

**Table2:** Categorical features’ statistical data

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Features** | **count** | **unique** | **top** | **freq** |
| checking\_balance | 1000 | 4 | unknown | 394 |
| credit\_history | 1000 | 5 | good | 530 |
| purpose | 1000 | 6 | furniture/appliances | 473 |
| savings\_balance | 1000 | 5 | < 100 DM | 603 |
| employment\_duration | 1000 | 5 | 1 - 4 years | 339 |
| other\_credit | 1000 | 3 | none | 814 |
| housing | 1000 | 3 | own | 713 |
| job | 1000 | 4 | skilled | 630 |
| phone | 1000 | 2 | no | 596 |
| default | 1000 | 2 | no | 700 |

## Appendix B: Exploratory Data Analysis – Visualization.

**Image1**: Numeric feature distribution

A diagram of a number of numbers

Description automatically generated with medium confidence

**Image2**: Numeric Features correlation.

A screenshot of a graph

Description automatically generated

## Appendix C: Model performance accuracy and ROC\_AUC score

**Table3:** Model performance comparison in term of train, test accuracy and ROC\_AUC score.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **model\_type** | **model** | **train** | **test** | **ROC\_AUC** |
| non\_tree\_base\_model | LogisticRegression | 0.721 | 0.713 | 0.754 |
| KNN | 0.697 | 0.697 | 0.663 |
| LDA | 0.717 | 0.710 | 0.751 |
| QDA | 0.687 | 0.697 | 0.734 |
| SVM | 0.724 | 0.740 | 0.749 |
| Voting | 0.790 | 0.727 | 0.748 |
| tree\_base\_model | Bagging | 0.766 | 0.753 | 0.756 |
| **RandomForest** | 0.757 | 0.743 | 0.775 |
| GradientBoosting | 0.756 | 0.720 | 0.772 |
| **AdaBoost** | 0.739 | 0.737 | 0.789 |
| XGB | 0.764 | 0.753 | 0.768 |
| LGBM | 0.766 | 0.723 | 0.761 |
| CatBoost | 0.767 | 0.743 | 0.768 |
| Voting | 0.967 | 0.730 | 0.779 |

**Image3**: Model type performance

A chart of different types of performance

Description automatically generated

**Image4**: Model performance comparison.

A graph of different types of performance

Description automatically generated with medium confidence

**Image5**: ROC\_AUC curve

A graph of a graph

Description automatically generated with medium confidence